FACULDADE DE TECNOLOGIA SÃO PAULO

PROCESSAMENTO DE DADOS

MONOGRAFIA

MODELAGEM DIMENSIONAL

SÃO PAULO

2012

FACULDADE DE TECNOLOGIA SÃO PAULO

PROCESSAMENTO DE DADOS

MONOGRAFIA

MODELAGEM DIMENSIONAL

RAMON RAMOS DE CASTRO NOVAIS

Monografia apresentada à Faculdade de Tecnologia São Paulo para obtenção de grau Tecnólogo em Processamento de Dados

Professor Orientador: Gabriel Shammas

SÃO PAULO

2012

DEDICATÓRIA

Aos meus pais

Rogério e Wanderly

Às minhas avós

Nair e Adelina

Aos meus amigos de Faculdade e Trabalho

(cujos nomes não serão citados para não correr o risco de esquecer alguém)

AGRADECIMENTOS

Primeiramente à minha namorada pelo apoio, ajuda e paciência durante o processo de elaboração deste trabalho.

Aos meus pais pelo apoio, tanto emocionalmente quanto financeiramente para minha conclusão da faculdade.

Ao meu professor orientador não só pela ajuda neste trabalho, más também pelos ensinamentos passados em sala de aula, que foram decisivos na escolha do tema e até minha carreira profissional.

E, finalmente a todos que contribuíram direta e indiretamente para realização deste trabalho.

SUMÁRIO

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE FIGURAS

RESUMO

Um data warehouse consiste em uma coleção de dados orientada por assuntos integrados, variante no tempo e não volátil que dá suporte à tomada de decisão pela alta gerência da empresa.

É também um conjunto de ferramentas e técnicas de projeto, que quando aplicadas às necessidades específicas dos usuários e aos bancos de dados específicos permitirá que planejem e construam um Depósito de Dados.

O Data Warehouse não é um produto e não pode ser comprado como um software de banco de dados. O sistema de Data Warehouse é similar ao desenvolvimento de um ERP, ou seja, ele exige análise do negócio, exige o entendimento do que se quer retirar das informações. Apesar de existirem produtos que fornecem uma gama de ferramentas para efetuar o Cleansing dos dados, a modelagem do banco e da apresentação dos dados, nada disso pode ser feito sem um elevado grau de análise e desenvolvimento.

O sistema de Data Warehouse não pode ser aprendido ou codificado como uma linguagem. Devido ao grande número de componentes e de etapas, um sistema de Data Warehouse suporta diversas linguagens e programações desde a extração dos dados até a apresentação dos mesmos.

ABSTRACT

A Data Warehouse consists in a integrated colection of subject oriented datas, that varies on time and is not volatile which gives suport to the decision taked by the company's high management.

It is also a series of tools and project tecniques, that when applied to the users' specifics needs and specifics data bases will allow to be planed and built a Data Warehouse.

The Data Warehouse is not a product and can not be bought as a data base software. The Data Warehouse system is similar to the development of a ERP, in other words, it demands a business analysis, demands the understainding of the business and demands what is needed to retrieve of the informations. Although the existence of products that provide a huge number of tools to perform the data Cleaning, the data base modeling and the data presentation, none of this can be done without a high level of analysis and development.

The Data Warehouse system can not be learned or codified as a language. Due to its greats numbers os components and steps, a Data Warehouse system suports many languages and programing since the data extraction to the presentation of then.

INTRODUÇÃO

O mundo dos negócios a muito utiliza informações como vantagem competitiva. Com o crescimento do mercado com a globalização, a computação se tornou fundamental para processar grandes volumes de dados. À medida que os anos avançaram, empresas começaram a se tornar corporações e os sistemas saíram das áreas operacionais para influenciar nas decisões de negócio. Nesse cenário percebeu-se o grande gargalo dos antigos meios de armazenamento: a questão de desempenho para obter informação de grande volume de dados.

E não somente obter as informações a tempo, outro problema era o formato. Com as constantes aquisições de empresas menores, as fontes de dados de sistemas pré-existentes não conseguiam se comunicar devido a diferentes padrões de suas arquiteturas.

A solução dos dois problemas é abordada nesse trabalho: ETL e Datawarehouse. De um lado, veremos mais detalhadamente a necessidade de implantação de datawarehouse em médias/grandes corporações e o surgimento das ferramentas de ETL para possibilitar essa centralização de dados.

1. BUSINESS INTELLIGENCE

Segundo artigo publicado por Orlando Augusto Nunes1, Business Intelligence é um conjunto de conceitos e metodologias que, fazendo uso de acontecimentos (fatos) e sistemas baseados nos mesmos, apoia a tomada de decisões em negócios. O grande desafio de todo indivíduo que gerencia qualquer processo é a análise dos fatos relacionados a seu dever. Ela deve ser feita de modo que, com as ferramentas e dados disponíveis, o gerente possa detectar tendências e tomar decisões eficientes e no tempo correto. Com essa necessidade surgiu então o conceito de Business Intelligence.

O propósito do Business Intelligence é permitir a tomada de decisões proativas, ao gerar informações necessárias ao negócio e disponibilizá-los no momento certo. À medida que o cenário econômico muda cada vez mais rápido, a necessidade de informações de negócios e as demandas pela rapidez e qualidade destas informações crescem na mesma velocidade. Por outro lado, a oferta de informações de negócios aumenta constantemente. O resultado é uma overdose de dados onde é difícil retirar informações relevantes para subsidiar tomadas de decisão. Tal fato torna mais difícil um entendimento aprofundado do cenário econômico. Faz-se necessária uma aproximação/abordagem sistemática para analisar temas e tendências estratégicas e antever mudanças com clientes, atividades e competidores.

1.1.História

1 Tecnólogo em Planejamento de Transportes pelo (IFG)

Há milhares de anos, Fenícios, Persas, Egípcios e outros Orientais já faziam, a seu modo, Business Intelligence, ou seja, cruzavam informações provenientes da natureza, tais como comportamento das marés, períodos de seca e de chuvas, posição dos astros, para tomar decisões que permitissem a melhoria de vida de suas comunidades. A história do Business Intelligence que conhecemos hoje, começa na década de 70, quando alguns produtos de BI foram disponibilizados para os analistas de negócio. O grande problema era que esses produtos exigiam intensa e exaustiva programação, não disponibilizavam informação em tempo hábil nem de forma flexível, e além de tudo tinham alto custo de implantação.

Em 1989, Howard Dresner2, um pesquisador da consultoria Gartner Group3, popularizou o BI como um termo geral para descrever um grupo de conceitos e métodos para melhorar o processo de tomada de decisão em negócios através do uso de sistemas.

Também no final dos anos 80 e início dos anos 90, surgiu o conceito de estocagem de informação. A ideia era deixar os dados onde eles já estavam e acessá-los a partir de qualquer lugar com alguma ferramenta específica. Com o aumento de padrões, automações e tecnologias, uma vasta quantidade de dados se tornou disponível. As tecnologias de estocagem de dados determinaram repositórios para armazenar estes dados. A melhora das ferramentas que extraiam, transformavam e carregavam dados aumentaram a velocidade de coleta de informações. Tecnologias permitiram a rápida geração de novos relatórios para a análise de dados, tudo pela ação do próprio usuário, sem interferência ou dependência de profissionais de tecnologia da informação (TI). O BI agora se tornava a arte de filtrar uma grande quantidade de dados, extrair as informações pertinentes, e tornar a informação em conhecimento.

2 Vice presidente da Gartner Inc. na época, deixou a empresa em 2005.

3 Consultoria fundada em 1979 por Gideon Gartner

1.2.Cenário Atual

Por muitos anos BI tem sido utilizado por grandes corporações apenas, já que essas eram as únicas capazes de investir grandes quantias, tanto de tempo quanto de dinheiro, para desenvolver os projetos de BI.

Atualmente uma nova geração de produtos de BI tem surgido: alta performance, operações intuitivas e rápida implantação são características das novas soluções que atendem a empresas de todos os tamanhos.

Essas novas soluções se mantém fieis ao objetivo inicial de apoiar a tomada de decisão transformando dados em conhecimento. No entanto, em vez de estar atado a questões de estrutura e conformidade que caracterizam os antigos produtos de BI, a nova abordagem foca na implantação simples e resultados rápidos. A importância dessa economia de tempo não pode ser subestimada já que o custo de projetos longos não está apenas no seu valor absoluto, mas no tempo que se gasta e na falta de resultados no tempo que o mercado necessita.

1.3.Datawarehouse

Datawarehouse pode ser definido com as duas palavras que compõe seu nome: Data, do inglês, dados; Warehouse, do inglês, armazém, local para estocar seus bens. Datawarehouse é um local onde é guardado toda informação estratégica para empresa para auxiliar na tomada de decisão.

Para obter uma melhor analise de quais dados serão guardados no datawarehouse, deve-se entender a diferença dos tipos de dados de um empresa:

- Dados operacionais

  Produzido por softwares corporativos, as empresas utilizam para serviços rotineiros como pedidos de clientes, cadastro de produtos ou gerenciamento de transações financeiras. Dados operacionais também podem vir de fontes externas como por exemplo sistemas de cotação financeira.

- Dados de integração

  Dados utilizados para integrar aplicações que não foram projetados para trabalhar juntas.

- Dados para controle

  Utilizado para elaboração de relatórios de apoio à decisão. Os dados são preparados para permitir aos usuários a melhor compreensão da situação da empresa.

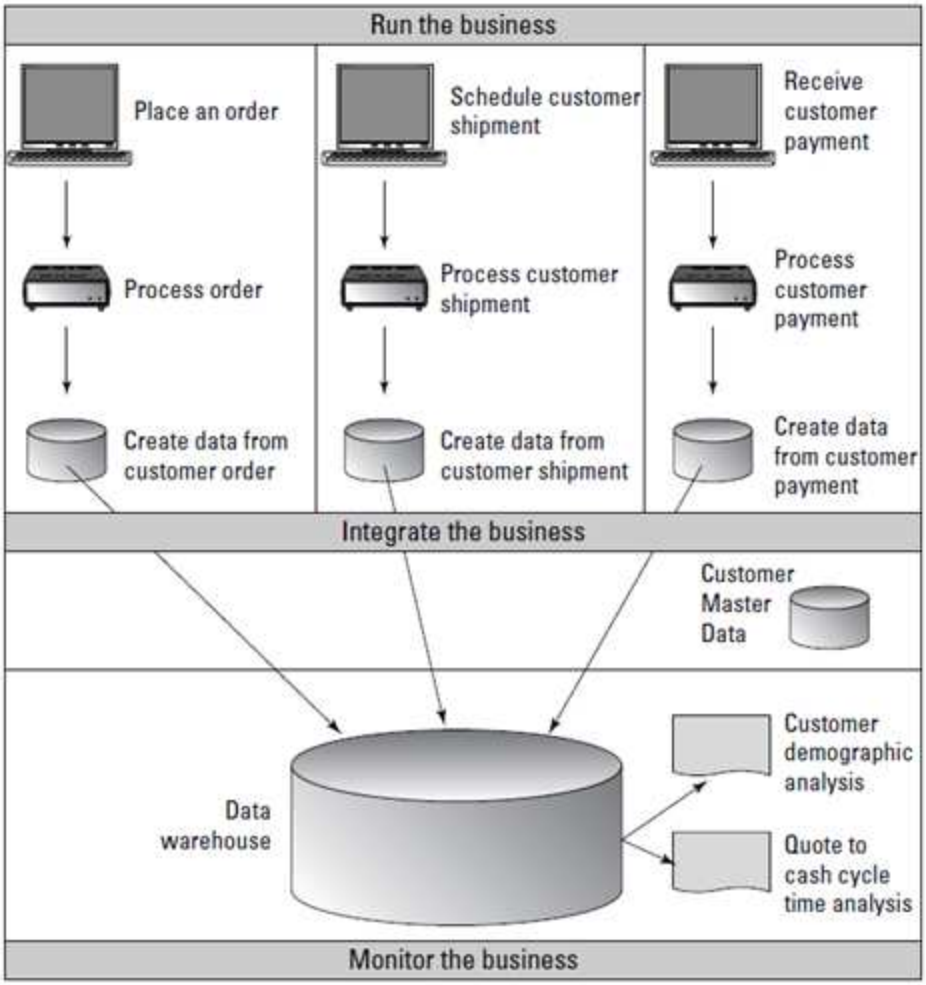

Figura 1 – Entrada de dados em data warehouse

Repare que na imagem acima o datawarehouse apenas recebe as informações de diversos sistemas. Os datawarehouses não têm como função gerar dados operacionais, mas tratar esses dados junto com os dados de integração para gerar dados de controle e melhoria corporativa.

2. INTRODUÇÃO A BANCO DE DADOS

Bancos de dados (ou bases de dados) são conjuntos de dados logicamente coerentes dispostos em estrutura regular que possibilita a reorganização dos mesmos e produção de informação. Representa abstratamente uma parte do mundo real que é de interesse de certa aplicação.

Por esta definição qualquer sistema que reúna e mantenha organizada uma série de informações relacionadas a um determinado assunto em uma determinada ordem pode ser considerado um banco de dados, como por exemplo, uma lista telefônica, um arquivo de clientes, uma lista de produtos, etc.

2.1.História do Banco de Dados

Nas décadas de 1960 e 1970, as empresas descobriram que estava muito custoso empregar um número grande de pessoas para fazer trabalhos como armazenar e indexar (organizar) arquivos. Por este motivo, valia a pena os esforços e investimentos em pesquisar um meio mais barato e ter uma solução mecânica eficiente.

Em 1970 um pesquisador da IBM - Ted Codd 4- publicou o primeiro artigo sobre bancos de dados relacionais. Este artigo tratava sobre o uso de cálculo e álgebra relacional para permitir que usuários não técnicos armazenassem e recuperassem grande quantidade de informações. Codd visionava um sistema onde o usuário seria capaz de acessar as informações através de comandos em inglês, onde as informações estariam armazenadas em tabelas.

4 Edgard Frank “Ted” Codd (1923-2003) cientista computacional inglês. Realizou várias contribuições à ciência computacional, mas o modelo relacional permaneceu como seu maior feito.

Devido à natureza técnica deste artigo e a relativa complicação matemática, o significado e proposição do artigo não foram prontamente realizados. Entretanto ele levou a IBM a montar um grupo de pesquisa conhecido como System R5 (Sistema R).

O projeto do Sistema R era criar um sistema de banco de dados relacional o qual eventualmente se tornaria um produto. Os primeiros protótipos foram utilizados por muitas organizações, tais como MIT Sloan School of Management6. Novas versões foram testadas com empresas aviação para rastreamento do manufaturamento de estoque.

Eventualmente o Sistema R evoluiu para SQL/DS7, o qual posteriormente tornou-se o DB28. A linguagem criada pelo grupo do Sistema R foi a Structured Query Language (SQL - Linguagem de Consulta Estruturada). Esta linguagem tornou-se um padrão na indústria para bancos de dados relacionais e hoje em dia é um padrão ISO 9 (International Organization for Standardization).

2.1.1. Os primeiros do mercado

Mesmo a IBM sendo a companhia que inventou o conceito original e o padrão SQL, eles não produziram o primeiro sistema comercial de banco de dados. O feito foi

5 Sistema de banco de dados constituído como um projeto da IBM San Jose Research (agora IBM Almaden Research Center) em 1970. Precursor do SQL, foi o primeiro sistema a demonstrar que um modelo relacional poderia oferecer um bom desempenho em processamento de transações.

6 Área de business do Instituto Tecnológico de Massachusetts, em Cambridge.

7 Structured Query Language/Data System. Uma implementação imperfeita do modelo relacional de Ted Codd, foi o primeiro uso comercial de um DBMS (Database Management System) da IBM em seus mainframes utilizando a linguagem SQL. Introduzido no mercado no inicio de 1980.

8 Criando em 1983 pela IBM. Existem diferentes versões do DB2 que rodam em desde um simples PDA, até em potentes mainframes e funcionam em servidores baseados em sistemas Unix, Windows, ou Linux.

9 Organização Internacional de Normalização/Padronização. Fundada em 1947, em Genebra, Suíça, a ISSO aprova normas internacionais em todos os campos técnicos.

realizado pela Honeywell Information Systems Inc10., cujo sistema foi lançado em junho de 1976. O sistema era baseado em muitos princípios do sistema que a IBM concebeu, mas foi modelado e implementado fora da IBM.

O primeiro sistema de banco de dados construído baseado nos padrões SQL começaram a aparecer no início dos anos 80 com a empresa Oracle através do Oracle 2 e depois com a IBM através do SQL/DS, servindo como sistema e repositório de informações de outras empresas.

2.1.2. Evolução

O software de banco de dados relacionais foi sendo refinado durante a década de 80. Isso deve-se ao feedback (retorno) que os usuários destes sistemas faziam, devido ao desenvolvimento de sistemas para novas indústrias e ao aumento do uso de computadores pessoais e sistemas distribuídos.

Desde sua chegada, os bancos de dados têm tido aumento nos dados de armazenamento, desde os 8 MB (Megabytes) até centenas de Terabytes de dados em listas de e-mail, informações sobre consumidores, sobre produtos, vídeos, informações geográficas, etc. Com este aumento de volume de dados, os sistemas de bancos de dados em operação também sofreram aumento em seu tamanho.

O padrão SQL passou da IBM para a ANSI 11 e para a ISO, os quais formaram um grupo de trabalho para continuar o desenvolvimento.

---

10 Grande conglomerado de empresas americanas seriada em Morristown, Nova Jersey. Produz uma variedade de produtos comerciais e de consumo, serviços de engenharia e sistemas aeroespaciais para uma ampla variedade de clientes, desde consumidores privados a grandes corporações e governos.

2.2. Banco de Dados Relacionais

A modelagem relacional baseia-se no princípio de que as informações em uma determinada base de dados podem ser consideradas como relações matemáticas e que estão representadas de forma uniforme, através do uso de tabelas bidimensionais. Desta forma, os dados são dirigidos para estruturas mais simples de armazenamento os dados, que são as tabelas, na qual a visão do usuário é privilegiada.

Este modelo, por suas características e por sua completitude, mostrou ser uma excelente opção, superando os modelos mais usados àquela época: o de redes 12e o hierárquico13. A maior vantagem do modelo relacional sobre seus antecessores é a representação simples dos dados e a facilidade com que consultas complexas podem ser expressas.

Esse modelo surgiu para atender sistemas transacionais que possuem operações atômicas (que devem ocorrer por completo ou então serão desfeitas) pré-definidas, geralmente com um grande número de usuários simultâneos realizando operações repetidamente.

2.2.1. Elementos

11 American National Standards Institute, com sede em Washington, DC. Organização privada sem fins lucrativos que supervisiona o desenvolvimento de padrões de consenso voluntário para produtos, serviços, processos, sistemas e pessoal nos Estados Unidos.

12 Neste modelo as entidades se representam como nós e suas relações são as linhas que os unem. Nesta estrutura qualquer componente pode se relacionar com qualquer outro como em uma teia de aranha.

13 Utiliza árvores para a representação lógica dos dados. Estas árvores são compostas de elementos chamados nós. O nível mais alto da árvore denomina-se raiz. Cada nó representa um registro com seus correspondentes campos.

- Tabela (ou Relações, ou Entidades): Armazena todos os dados do BDR (Banco de Dados Relacional), e é composta de linhas e colunas. Essas tabelas associam-se entre si através de regras de relacionamento, estas regras consistem em associar um atributo de uma tabela com um conjunto de registros de outra tabela.
- Registros (ou Tuplas): Cada linha de uma tabela representa um registro (ou tupla); um registro é uma instância de uma tabela, ou entidade. Esses registros não precisam necessariamente conter informações em todas as colunas, podendo admitir valores nulos quando necessário.
- Atributo: colunas de uma tabela
- Chave: conjunto de um ou mais atributos que determinam a unicidade de cada registro

3. BANCO DE DADOS DIMENSIONAL

Diferente dos modelos anteriores (hierárquico, de rede e relacional) onde as informações eram divididas em muitas tabelas para melhor representação do mundo físico, o foco no modelo dimensional é agrupamento de informação. Onde nos seus predecessores redundância de dados era visto como um defeito, aqui se torna uma vantagem competitiva. Não é a representação computacional do mundo físico que se busca nessa abordagem, mas a coletânea de informação para alimentar sistemas necessariamente ágeis. Aqui o cliente não está interessado em ver o modelo lógico, mas nos números finais de sua consulta, as vezes gráficos ou tabelas, que tenham potencial para influenciar nas suas decisões de negócio.

3.1. História

1970, a preparação:

Década tecnológica com predominância de mainframes. Apesar do desempenho em executar funções rotineiras, os dados criados desse processamento são isolados em bancos de dados primitivos e conjunto de arquivos apenas acessíveis aos departamentos de processamento de dados responsáveis pelo mainframe.

Era quase impossível, por exemplo, comparar o desempenho de lojas de varejo da região oriental de um grupo de empresas com as lojas da região ocidental, ou com seus concorrentes ou mesmo contra seu próprio desempenho em um período anterior.

Como tentativa de obter essas informações, grandes demandas de relatórios eram frequentes aos departamentos de processamento de dados, gerando filas constantes de pendências.

A fim de suprir essas necessidades em tempo mais hábil, algumas empresas adotaram uma abordagem interessante: eram identificados dados importantes e constantemente requisitados (como informações de clientes, vendas e despesas) e periodicamente era feito a copia dos dados para fonte externa onde esses dados poderiam ser acessados para formar relatórios comuns (como relatórios de lucro, despesas, ganho por cliente).

O problema dessa abordagem era sua aplicabilidade. Em vista que a fonte inicial dos dados eram mainframes de certas empresas, comparativos entre empresas com mainframes configurados de forma diferente já não era possível.

Empresas de hardware/software começaram a surgir com soluções para esse problema. Entre 1976 e 1979, a partir da pesquisa do instituto tecnológico da Califórnia (Caltech14) e o grupo de tecnologias avançadas do Citybank surgiria a empresa Teradata. Seus fundadores desenvolveram um sistema de gerenciamento de banco de dados para processamento paralelo utilizando vários microprocessadores focado em sistemas de suporte a decisão. Teradata 15foi constituída em 13 de julho de 1979 e começou em uma garagem em Brentwood, Califórnia. O nome do Teradata foi escolhido para simbolizar a habilidade de gerenciar terabytes (trilhões de bytes) de dados.

3.2. Introdução e conceito

De acordo com Ralph Kimball, “Dimensional Modeling is a design technique for databases intended to support end-user queries in a data warehouse.”, modelagem dimensional é uma técnica de design de banco de dados projetada para suportar consultas de end-users em um Data Warehouse. Para sistemas de processamento analítico, o grande volume de dados necessários para consultas de planejamento tático e estratégico devem ser processados de forma rápida. Para melhorar o desempenho, há redundância planejada dos dados (diferente do modelo relacional), compensando os gastos com armazenamento e atualização das informações. Com isso temos uma estrutura simples, com tabelas de dados históricos em series temporais, descritos através

14 Califórnia Institute of Technology. Universidade privada localizada em Pasadena, Califórnia. Sendo uma das primeiras universidades do mundo em pesquisa, a Caltech mantém uma forte ênfase e tradição nas Ciências naturais e Engenharia. De acordo com a classificação anual da Times Higher Education de 2011, a Caltech é a melhor universidade do mundo.

15 Teradata Corporation – fundada em 1979. Considerada pela pesquisa do Instituto Gartner uma das empresas lideres em datawarehouse e ferramentas business analytics.

de tabelas de dimensões, de modo que o modelo reflita o processo de analise de negócios da empresa.

As atualizações nesse modelo são feitas periodicamente em batch, não havendo necessidade de controlar as alterações realizadas entre uma atualização e outra.

Um dos objetivos desse modelo é permitir ao usuário realizar consultas na base de dados sem depender da equipe de tecnologia.

As diferenças entre os modelos relacionais e dimensionais pode ser visto na tabela a seguir:

Processamento Transacional	Processamento Analítico
Dados Normalizados	Dados Consistentes
Atualização em tempo real	Desempenho compatível com o volume de dados
Controle de Concorrências	Cala representação do modelo
Dados Concorrentes	Ferramentas especiais para usuários finais
Respostas imediatas	

Tabela 1- Comparativo entre processamento transacional e analítico

3.3. Modelagem dimensional

Elementos que compõem um Modelo Dimensional:

- Tabela Fato

De acordo com Kimball, a tabela de fatos é a principal tabela de um modelo dimensional, onde as medições numéricas de interesse da empresa estão armazenadas. A palavra "fato" representa uma medida dos processos modelados, como quantidades, valores e indicadores. A tabela de fatos registra os fatos que serão analisados. É composta por uma chave primária (formada por uma combinação única de valores de chaves de dimensão) e pelas métricas de interesse para o negócio.

A tabela de fatos deve ser sempre preenchida com as medidas referentes ao fato. Não se deve preencher uma linha da tabela fato com zeros para representar que nada aconteceu (por exemplo, que não houve vendas de um produto em determinada data), pois isso faria com que a tabela de fatos crescesse demais. Além disso, a tabela de fatos deve representar uma unidade do processo do negócio, não devendo-se misturar assuntos diferentes numa mesma tabela de fatos.

- Tabela Dimensão

A tabela de dimensão é composta de atributos e contém a descrição do negócio. Seus atributos são fontes das restrições de consultas, agrupamento dos resultados e cabeçalhos para relatórios. Ela possui aspectos pelos quais se pretende observar as métricas relativas ao processo modelado. A tabela de dimensão costuma ser bem menor do que a tabela fato, geralmente com muito menos do que um milhão de registro.

Um exemplo da diferença entre a tabela fato e a dimensão está na figura a seguir:

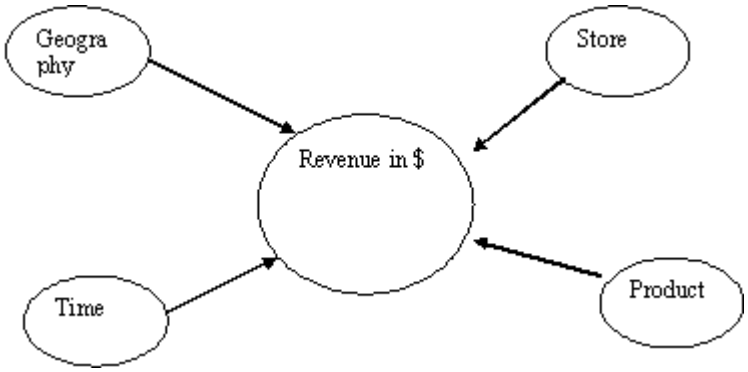

Gráfico 1 - Tabela fato no centro e demais tabelas dimensão ao redor

Se o fato/métrica a ser medido/a for a receita de uma rede de supermercados, as dimensões para a avaliação da métrica seriam, por exemplo, a quantidade de lojas, a localização, os produtos vendidos e o tempo.

- Tabela Agregada

A tabela agregada é criada com dados da tabela fato, alterando sua granularidade, ou seja, ela sumariza os dados, gerando uma tabela menor. A tabela agregada é utilizada para otimizar o tempo de acesso de uma consulta ao banco de dados. É importante avaliar bem o ambiente para definir quais agregações devem ser criadas; a utilização das mesmas requer um esforço adicional de manutenção, além de aumentar o gasto com armazenamento, por isso deve-se sempre tentar criar tabelas agregadas que atendam a múltiplas consultas. Além disso, as tabelas agregadas podem ser temporárias; desta forma, deve-se levar em conta a possível extinção dessa tabela e os futuros efeitos causados devido a sua exclusão.

- Métricas

São as informações armazenadas nas tabelas fato que permite medir o desempenho dos processos do negócio. As métricas são geralmente volumétricas,

numéricas, podem ou não ser agregadas e na maioria das vezes são do tipo aditivas, ou seja, permitem operações como adição, subtração e médias. Existem também outros dois tipos de métricas, as métricas não aditivas e as semi-aditivas. As métricas não aditivas não podem ser manipuladas livremente, como valores percentuais ou relativos. Já as métricas semi-aditivas são valores que não podem ser somados em todas as dimensões.

3.4. Modelos de Implementação

3.4.1. Modelo Estrela (Star Schema)

O nome “estrela” se dá devido à disposição em que se encontram as tabelas, sendo a tabela fato centralizada relacionando-se com diversas outras tabelas de dimensão. Veja um exemplo da estrutura do Star Schema na figura a seguir:

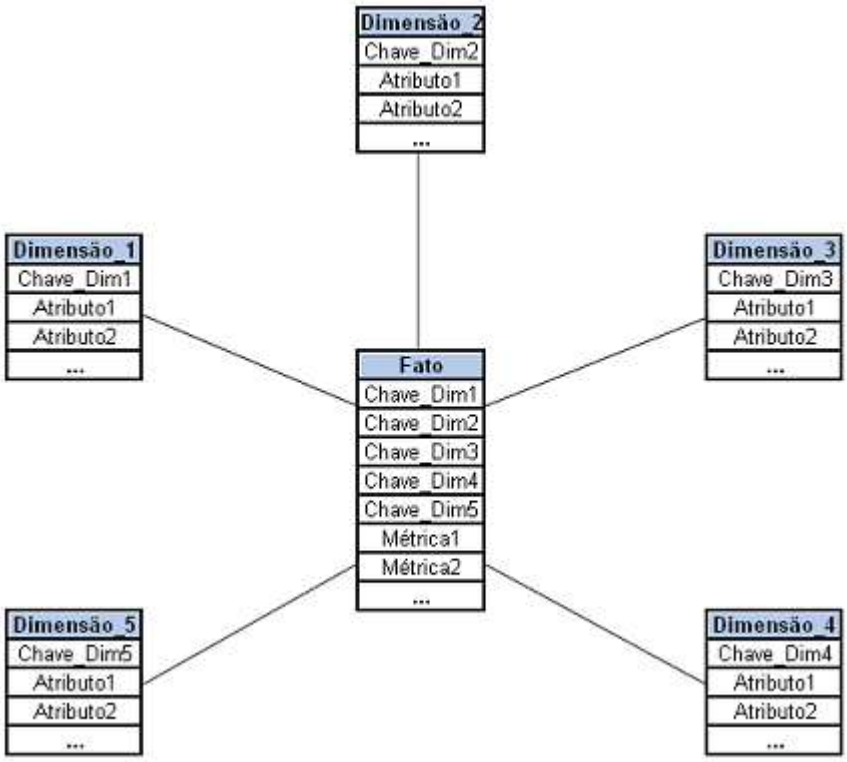

Gráfico 2 - Modelo Estrela

Nesse modelo os dados são desnormalisados para evitar joins entre tabelas, diminuindo o tempo de consultas, no entanto devido a repetição de dados, utiliza mais espaço em disco. A vantagem desse modelo é a eficiência na extração de dados, o que é um grande diferencial em se tratando de um datawarehouse.

3.4.2. Modelo Floco de Neve (Snow Flake)

Outro tipo de estrutura bastante comum, é o modelo de dados Snow Flake (Floco de Neve), que consiste em uma extensão do modelo Estrela onde cada uma das "pontas da estrela" passa a ser o centro de outras estrelas. Isto porque cada tabela de dimensão seria normalizada, "quebrando-se" a tabela original ao longo de hierarquias existentes

em seus atributos. Recomenda-se utilizar o esquema floco de neve apenas quando a linha de dimensão ficar muito longa e começar a ser relevante do ponto de vista de armazenamento. Veja um exemplo da estrutura do Snow Flakes na figura a seguir:

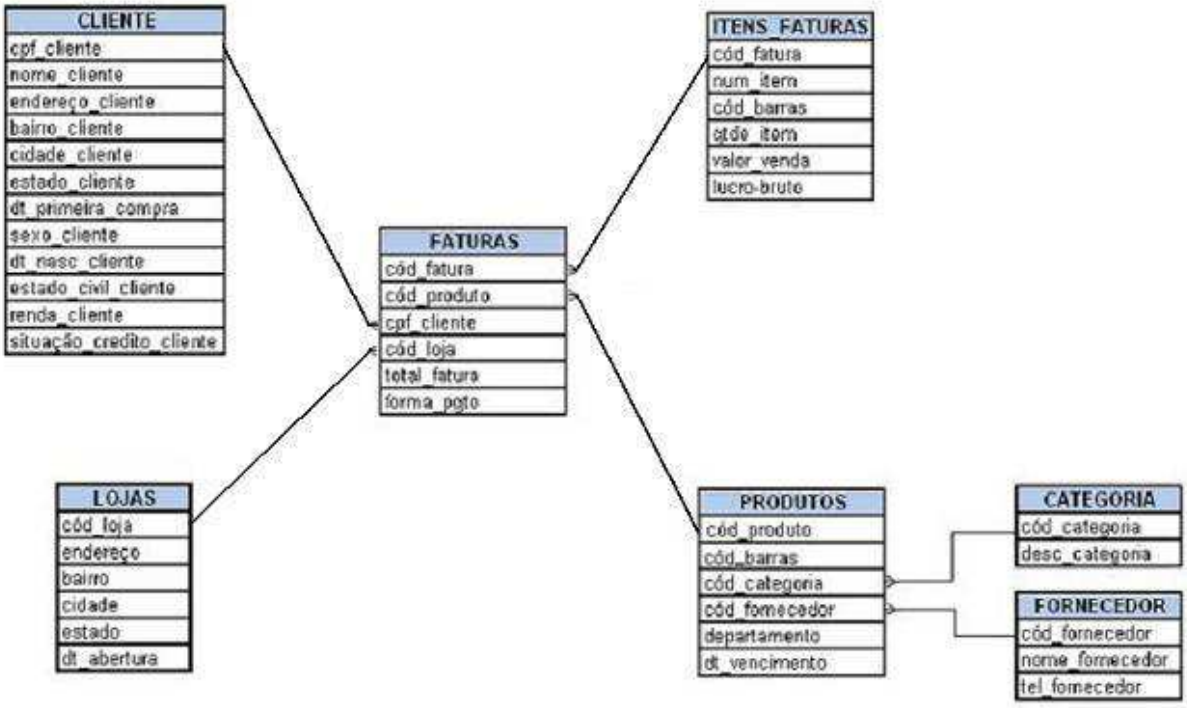

Gráfico 3 - Modelo Floco de neve

Devido a essa estrutura, o acesso aos dados é mais lenta, mas facilita na construção de cubos de algumas ferramentas BI(Business Intelligence) e BA (Business Analytics).

A decisão de optar pelo esquema estrela ou pelo floco de neve deve ser tomada levando-se em consideração o volume de dados, o SGBD, as ferramentas utilizadas, etc.

Abaixo temos uma tabela comparativa entre esses dois modelos.

	Modelo Estrela	Modelo Floco de Neve
Tabela dimensão	Não normalizada	Normalizada

Tamanho fisico	Grande volume já que os dados se repetem nas tabelas dimensões não normalizadas	Volume reduzido, já que os dados das tabelas dimensões são normalizados para evitar repetições
Velocidade das consultas	Rápida	Menos rápida do que o modelo estrela devido a normalização

Tabela 2 - Comparativo entre modelo estrela e floco de neve

4. ETL

ETL, vem do inglês Extract Transform Load, ou seja, Extração Transformação Carga. O ETL visa trabalhar com toda a parte de extração de dados de fontes externas, transformação para atender às necessidades de negócios e carga dos dados dentro do Data Warehouse.

Os projetos de data warehouse consolidam dados de diferentes fontes. A maioria dessas fontes tendem a ser bancos de dados relacionais ou flat files16, mas podem existir outros tipos de fontes também. Um sistema ETL precisa ser capaz de se comunicar com bases de dados e ler diversos formatos de arquivos utilizados por toda a organização.

16 Coleção de dados armazenados e acessados de forma sequencial que contem registros sem relação estruturada

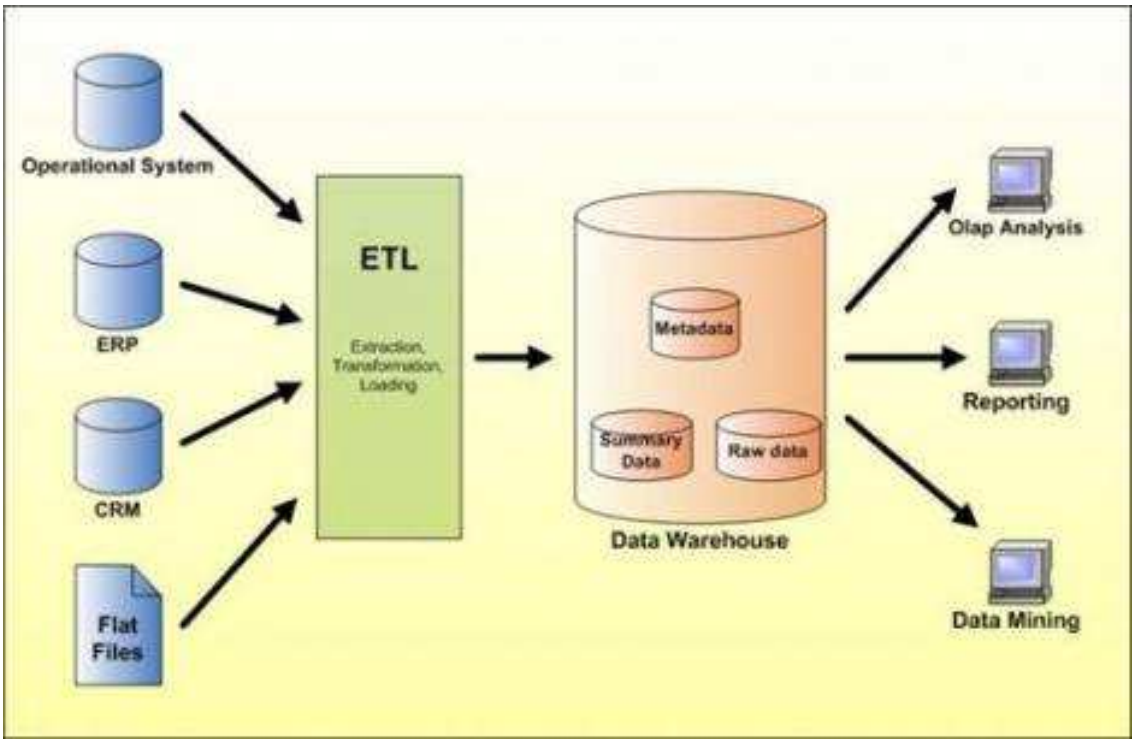

Gráfico 4 - ETL no ciclo de dados

Na imagem acima temos as entradas de dados em azul. A abordagem aqui é a possibilidade de diferentes tipos de entrada. Como as ferramentas de ETL unificam os dados, não há mais a necessidade de ficar preso a alguma marca ou tipo de armazenamento de dados. Podemos por exemplo ter os dados de um ERP, de uma ferramenta CRM, flat files ou bancos de dados de sistemas operacionais como RDBMS (Relational Data Base Management System, sistema gerenciador de banco de dados relacional) e mainframes, todos usando bases diferentes e podendo estar normalizados diferentemente também. Isto dá às empresas liberdade na escolha de suas ferramentas, visto que o importante agora são as vantagens operacionais e não as restrições funcionais.

Ao passar pelo processo de ETL (quadrado verde), os dados são armazenados no datawarehouse de forma que seja possível recuperar:

- Raw data – os dados. Ex.: data de nascimento
- Metadata – os dados sobre os dados. Ex.: tipo do dado (se texto ou datetime)
- Summary data – o agrupamento dos dados, resumo. Ex.: média das idades

A partir desse ponto, o datawarehouse já está preparado para alimentar os sistemas de apoio a decisão como ferramentas OLAP, relatórios e Data Marts.

4.1. Componentes do ETL

Nesta imagem nós podemos visualizar um exemplo de modelo de Arquitetura de uma solução de BI. O objetivo aqui não é discutir sobre toda a arquitetura, mas visualizar os principais componentes que fazem parte de um sistema ETL.

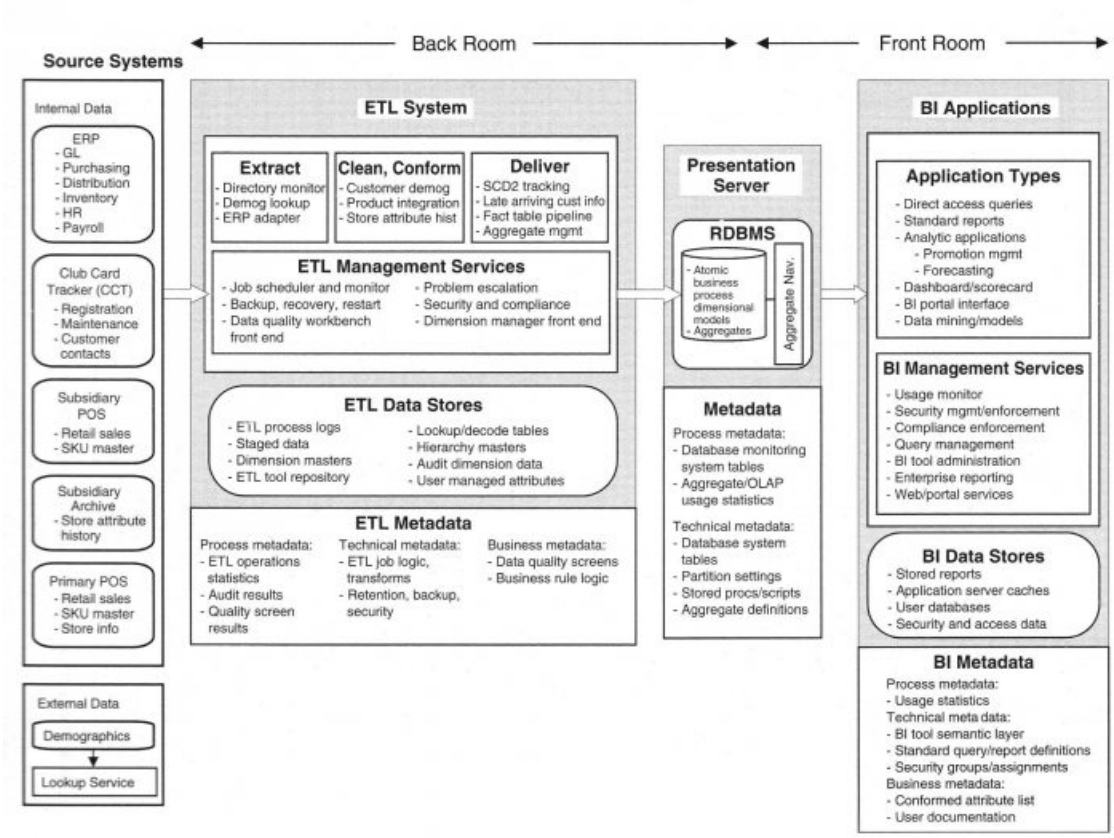

Figura 2 - Componentes do ETL

- Extração: É a coleta de dados dos sistemas de origem (também chamados Data Sources ou sistemas operacionais), extraindo-os e transferindo-os para o ambiente de DW, onde o sistema de ETL pode operar independente dos sistemas operacionais.
- Limpeza, Ajustes e Consolidação (ou também chamada transformação): É nesta etapa que realizamos os devidos ajustes, podendo assim melhorar a qualidade dos dados e consolidar dados de duas ou mais fontes.

O estágio de transformação aplica uma série de regras ou funções aos dados extraídos para ajustar os dados a serem carregados. Algumas fontes de dados necessitarão de muito pouca manipulação de dados. Em outros casos, pode ser necessários trabalhar algumas transformações, como por exemplo, junção de dados

provenientes de diversas fontes, seleção de apenas determinadas colunas e tradução de valores codificados (se o sistema de origem armazena 1 para sexo masculino e 2 para feminino, mas o data warehouse armazena M para masculino e F para feminino, por exemplo).

- Entrega ou Carga dos dados: Consiste em fisicamente estruturar e carregar os dados para dentro da camada de apresentação seguindo o modelo dimensional. Dependendo das necessidades da organização, este processo varia amplamente. Alguns data warehouses podem substituir as informações existentes semanalmente, com dados cumulativos e atualizados, ao passo que outro DW (ou até mesmo outras partes do mesmo DW) podem adicionar dados a cada hora. A latência e o alcance de reposição ou acréscimo constituem opções de projeto estratégicas que dependem do tempo disponível e das necessidades de negócios.
- A parte de Gerenciamento é composta por serviços para auxiliar no gerenciamento do DataWarehouse. Aqui nós temos tasks específicas para gerenciamento de jobs, planos de backup, verificação de itens de segurança e compliance.

4.2. ETL no Ciclo de Vida do Data Warehouse

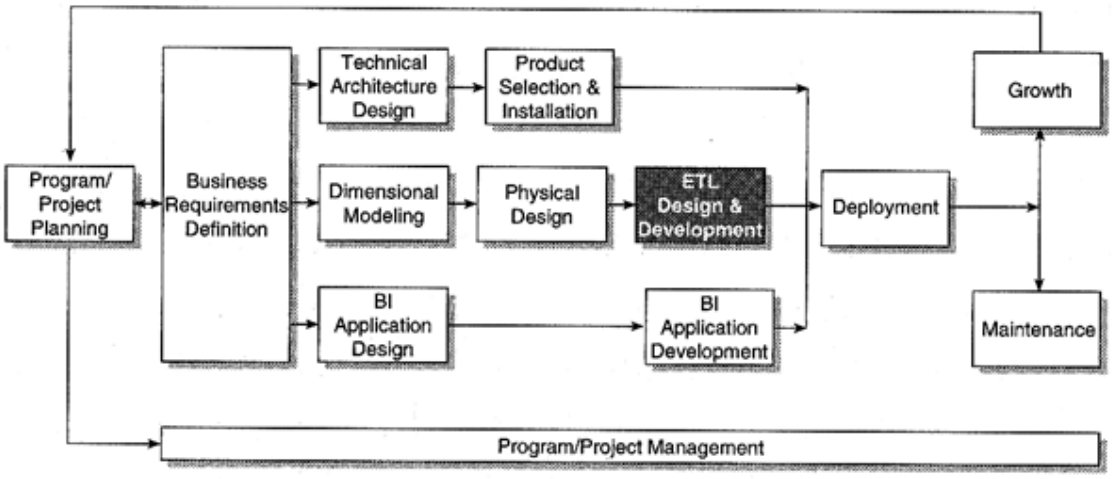

Gráfico 5 - ETL no ciclo de vida do data warehouse

O Ciclo de vida do Data Warehouse é composto por uma série de etapas. Inicia-se pelo planejamento do Programa ou Projeto, passamos pelo levantamento e definição dos requisitos de negócios e aí nos dividimos em três caminhos:

- Arquitetura e Design Técnico
- Modelagem Dimensiona
- Planejamento e desenvolvimento da aplicação de BI, o front-end17 propriamente dito.

5. DISTRIBUIÇÃO DE DATA WAREHOUSE

A maioria das organizações constrói e mantém um único data warehouse centralizado, isto é feito por várias razões:

Os dados em um data warehouse é integrado pela organização, e uma visão integrada dos dados é usada somente na sede da organização;

17 parte do sistema de software que interage diretamente com o usuário

A organização opera em um modelo centralizado de negócio;

- O volume dos dados em um data warehouse é tal que um único repositório de dados centralizado faz sentido;
- Complexidade de desenvolvimento de um ambiente distribuído;
- Maior Segurança; e
- Fácil Gerenciamento.

Em resumo, a política, a economia e a tecnologia favorecem muito o uso de um único data warehouse centralizado. Entretanto, dados extremamente centralizados podem resultar em perda de disponibilidade e queda de desempenho das consultas. Daí surge a necessidade de um ambiente de distribuição de data warehouse, tendo como vantagens sobre os ambientes centralizados: o aumento da disponibilidade dos dados, o aumento da disponibilidade de acesso aos dados e o aumento de desempenho no processamento de consultas OLAP.

5.1. Banco de Dados Distribuídos

Os banco de dados distribuídos trazem vantagens da computação distribuída para o domínio do gerenciamento de banco de dados. Um sistema de computação distribuída consiste em vários elementos de processamento, não necessariamente homogêneos, que são interconectados por uma rede de computadores e cooperam na execução de certas tarefas. Os banco de dados distribuídos podem ser definidos como uma coleção de múltiplos bancos de dados logicamente inter-relacionados, distribuídos

por uma rede de computadores. Abaixo são destacadas algumas vantagens na utilização de banco de dados distribuídos:

- Transparência de fragmentação, replicação e alocação;
- Melhoria na confiabilidade e disponibilidade;
- Melhoria de desempenho; e
- Expansão mais fácil;

De acordo com Elmasri e Navathe, a distribuição leva a uma maior complexidade no projeto e na implementação do sistema. Para obter as vantagens potenciais listadas anteriormente, o ambiente de banco de dados distribuídos deve ser capaz de prover algumas funções, além daquelas já presentes em ambientes centralizados, como por exemplo:

- Rastreamento dos dados;
- Processamento de consultas distribuídas;
- Gerenciamento de transações distribuídas;
- Gerenciamento de dados replicados;
- Recuperação de banco de dados distribuído;
- Segurança; e
- Gerenciamento do diretório (catálogo) distribuído.

5.1.1. Banco de Dados Distribuído x Data Warehouse Distribuído

O data warehouse nada mais é do que um banco de dados especial integrado, orientado por assunto, variável com o tempo e não volátil, usado para dar suporte ao

processo gerencial de tomada de decisão. Por isso, as contribuições obtidas pelos trabalhos de pesquisa em sistemas de banco de dados distribuídos podem ser utilizadas como base para o desenvolvimento de ambientes de data warehousing distribuídos. Porém, esses trabalhos devem ser estendidos de forma a enfocar aspectos importantes dos ambientes de data warehousing distribuído, tais como a multidimensionalidade dos dados do data warehouse, a organização dos dados dessa base de dados em diferentes níveis de agregação e as características das consultas OLAP comumente realizadas pelos usuários de sistemas de suporte à decisão.

5.2. Arquitetura de Data Warehouse Distribuído de Inmon

A arquitetura de data warehouse distribuído definida por Inmon 18é baseada

nos conceitos de data warehouse local e de data warehouse global. O Gráfico 6 ilustra esta arquitetura, onde o data warehouse global situa-se localizado no site correspondente ao escritório central ou sede da empresa, enquanto os data warehouses locais estão localizados em regiões geográficas diferentes ou comunidades técnicas distintas.

18 William H. Inmon (1945) – cientista computacional conhecido como pai do datawarehouse, criou a definição mais aceitável de datawarehouse: coleção de dados orientados, não voláteis, integrados, e variados pelo tempo para o auxiliar no suporte de decisões

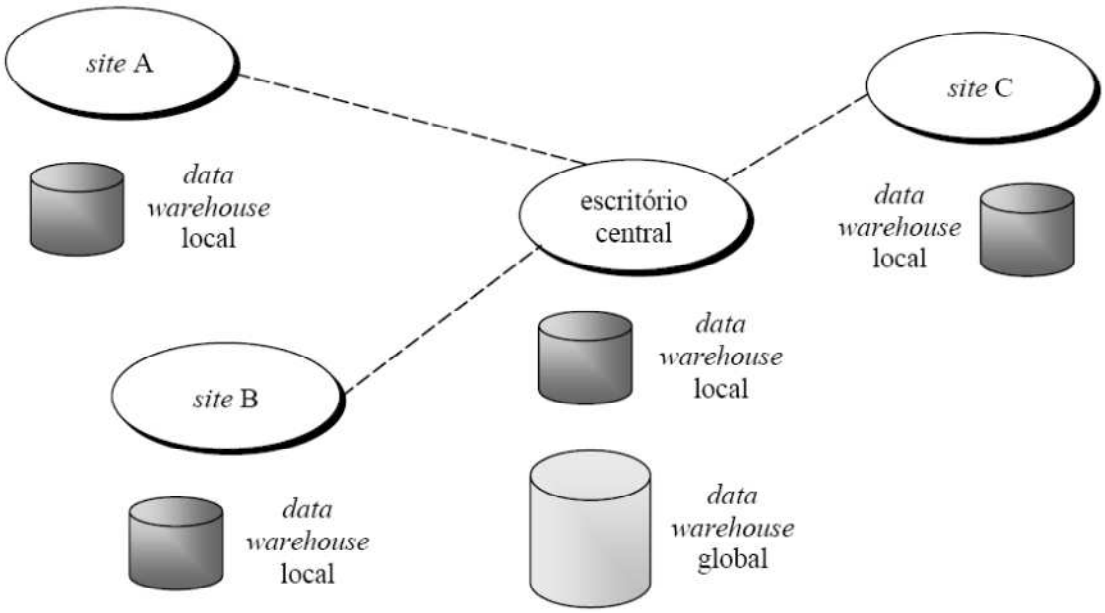

Gráfico 6 - Arquitetura básica do data warehouse distribuido de INMON

Os dados armazenados no data warehouse local são de interesse somente para o nível local, ou seja, cada data warehouse local tem como escopo dos seus dados detalhados que refletem a integração das informações provenientes dos sistemas operacionais do site local ao qual ele serve. Apesar de ser inteiramente possível a existência de algum grau de compartilhamento entre os sistemas do ambiente operacional encontrados em cada um dos sites, qualquer interseção ou compartilhamento dos dados de um data warehouse local para outro é apenas uma coincidência. Os dados armazenados no data warehouse global são de interesse para a empresa como um todo. Estes dados são integrados a partir das interseções naturais dos dados existentes nos sites que compõem o ambiente distribuído. O relacionamento entre o data warehouse global e os data warehouses locais pode ser observado da seguinte forma. Os dados levemente agregados residem no nível global, enquanto que os dados detalhados residem nos níveis locais. Como pode ser observado, os dados localizados no data warehouse global e nos data warehouses locais são mutuamente exclusivos:

qualquer dado no data warehouse global não é encontrado nos data warehouses locais, e vice-versa. Em contrapartida, o projeto estrutural dos dados corporativos armazenados no data warehouse global pode sobrepor porções dos modelos de dados dos data warehouses locais. Inmon propõe uma variação desta arquitetura, onde consiste no pré-armazenamento dos dados a serem enviados ao data warehouse global por cada um dos sites locais. Assim, cada site que participa do ambiente armazena os dados globais correspondentes às informações locais em uma base de dados especial, chamada de área de armazenamento do data warehouse global, antes de enviá-los ao data warehouse global propriamente dito. Neste caso, a restrição de exclusividade mútua dos dados é observada tanto entre os dados localizados nos data warehouses locais e nas áreas de armazenamento do data warehouse global quanto entre os dados localizados nos data warehouses locais e no data warehouse global. Contudo, pode haver alguma redundância entre os dados armazenados no data warehouse global e nas áreas de armazenamento do data warehouse global em cada um dos sites, caso a política adotada pela empresa seja a não remoção dos dados destas áreas após o envio destes ao data warehouse global. O Gráfico 7 representa as áreas de armazenamento do data warehouse global em cada um dos sites.

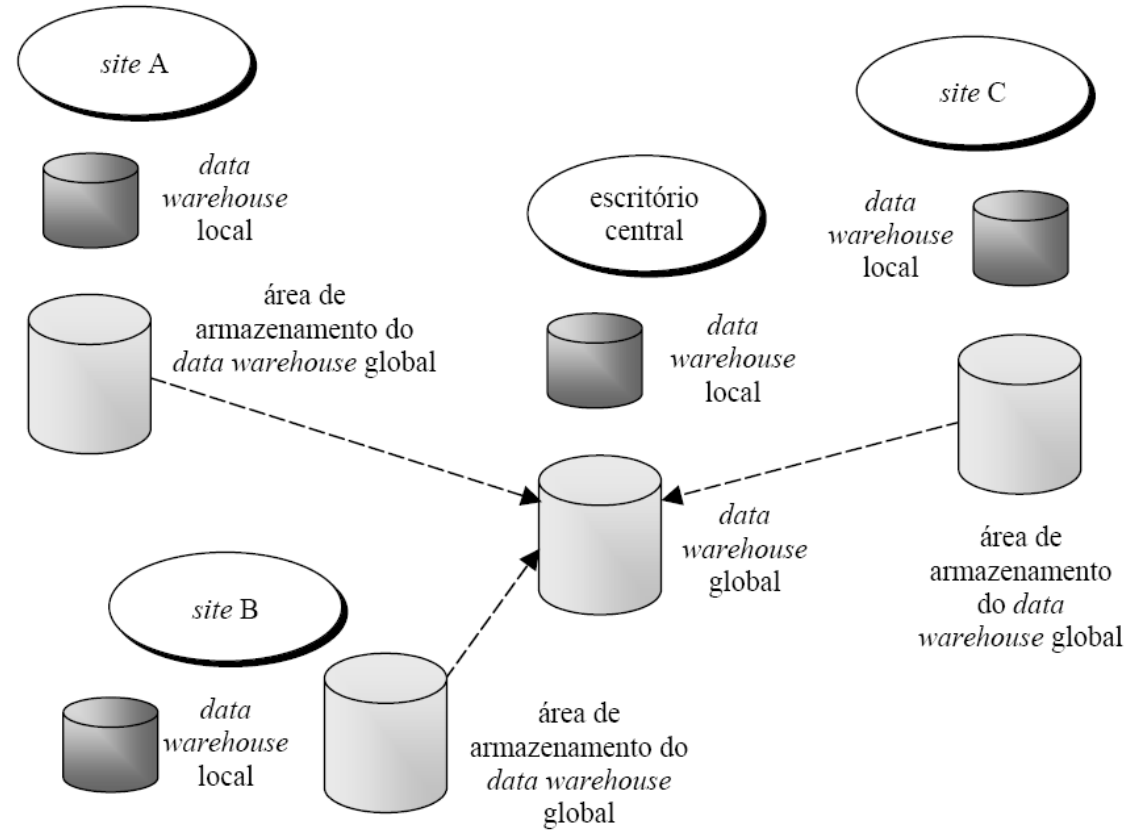

Gráfico 7 - Variação da arquitetura básica do data warehouse distribuído de INMON

Inmon sugere que o desenvolvimento desta arquitetura deve ser feito primeiramente criando os data warehouses locais para cada entidade geográfica, para que depois, o data warehouse global seja criado, refletindo a integração dos negócios através das diferentes localizações.

5.3. Arquitetura de Data Warehousing Distribuído de Moeller

As arquiteturas de data warehousing distribuído definidas por Moeller são baseadas na junção de dois conceitos: integração através do elemento banco de dados e distribuição através do elemento rede. Assim, um data warehouse distribuído é definido por Moeller como uma coleção de dados compartilhados logicamente integrada, a qual é

fisicamente distribuída através dos nós de uma rede de computadores. Uma vez que o data warehouse distribuído consiste na integração lógica de diversos bancos de dados locais, ele não existe fisicamente nas arquiteturas de Moeller. Mais especificamente, o data warehouse distribuído é apenas um conceito virtual. Em particular, os termos local e global são utilizados para realizar a distinção, respectivamente, entre os aspectos relacionados a um único site e os aspectos relacionados ao ambiente de data warehousing como um todo. Por exemplo, um data warehouse local refere-se a um banco de dados pré-existente que reside em um site específico da rede, ou seja, refere-se a um data mart.

Há três diferentes tipos de arquitetura de data warehousing distribuído apresentadas por Moeller [MOE01]: arquitetura de data warehousing distribuído homogêneo, heterogêneo e com um SGBD distribuído único.. É importante salientar que Moeller associa os seus três tipos de arquitetura de data warehousing distribuído à abordagem de desenvolvimento, na qual uma corporação já gerencia vários data marts independentes e deseja possibilitar, como uma atividade subseqüente, o acesso global dos usuários de SSD a estes data marts através de um data warehouse global virtual. Ou seja, os dados são mantidos nas fontes de dados e as consultas são decompostas em tempo real e submetidas às diversas fontes, onde o resultado é integrado e mostrado para o usuário que efetuou a consulta. Isto é obtido através do desenvolvimento de um esquema global da empresa como um todo, que representa a integração dos esquemas locais dos data marts existentes, além da interconexão destes data marts através da rede.

5.3.1. Arquitetura de Data Warehousing Distribuído Homogêneo

O Gráfico 8 mostra a arquitetura de data warehousing distribuído homogêneo proposta por Moeller, com os seus dois principais componentes: o data warehouse distribuído e a ferramenta de banco de dados distribuídos.

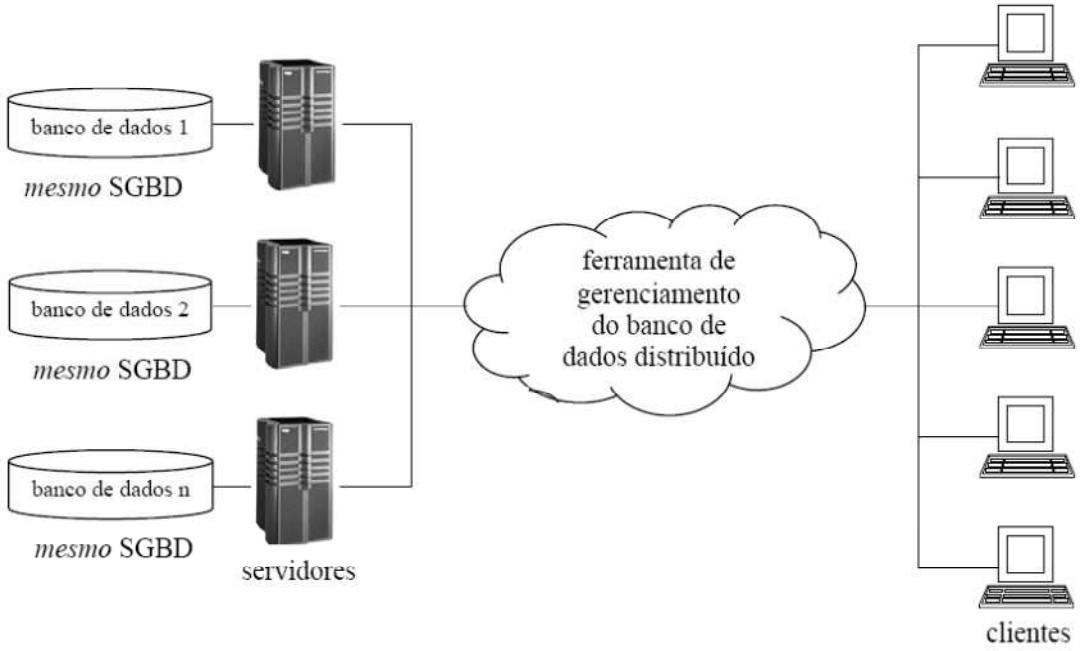

Gráfico 8 - Arquitura do data warehouse homogênio de MOELLER

Cada site nesta arquitetura possui o seu próprio banco de dados autônomo e pode representar um data mart independente. A arquitetura homogênea é caracterizada por apresentar em todos os sites o mesmo SGBD. São nestes SGBD que se armazenam os data marts a serem distribuídos. A ferramenta de gerenciamento do banco de dados distribuído, por sua vez, é responsável por integrar os diversos bancos de dados locais, oferecendo uma visão lógica do data warehouse corporativo, além de gerenciar as consultas dos usuários de SSD aos bancos de dados fora de suas redes locais. Essa ferramenta é baseada em dois elementos centrais relacionados à manipulação dos dados distribuídos: esquema de fragmentação e esquema de alocação. O esquema de fragmentação descreve como os relacionamentos globais são divididos entre os bancos de dados locais. Já o esquema de alocação especifica a localização de cada um dos

fragmentos, possibilitando a execução de consultas através dos diversos bancos de dados locais. Este último esquema também dá suporte à possibilidade de replicação dos dados na arquitetura.

5.3.2. Arquitetura de Data Warehousing Distribuído Heterogêneo

A arquitetura de data warehousing distribuído heterogêneo proposta por Moeller é baseada nos mesmos componentes principais que a arquitetura de data warehousing distribuído homogêneo: o data warehouse distribuído e uma ferramenta de gerenciamento do banco de dados distribuído. No entanto, na arquitetura de data warehousing distribuído heterogêneo, estes componentes possuem características e funcionalidades particulares relacionadas à heterogeneidade dos dados, aumentando, com isso, a complexidade destes componentes. O Gráfico 9 ilustra esta arquitetura.

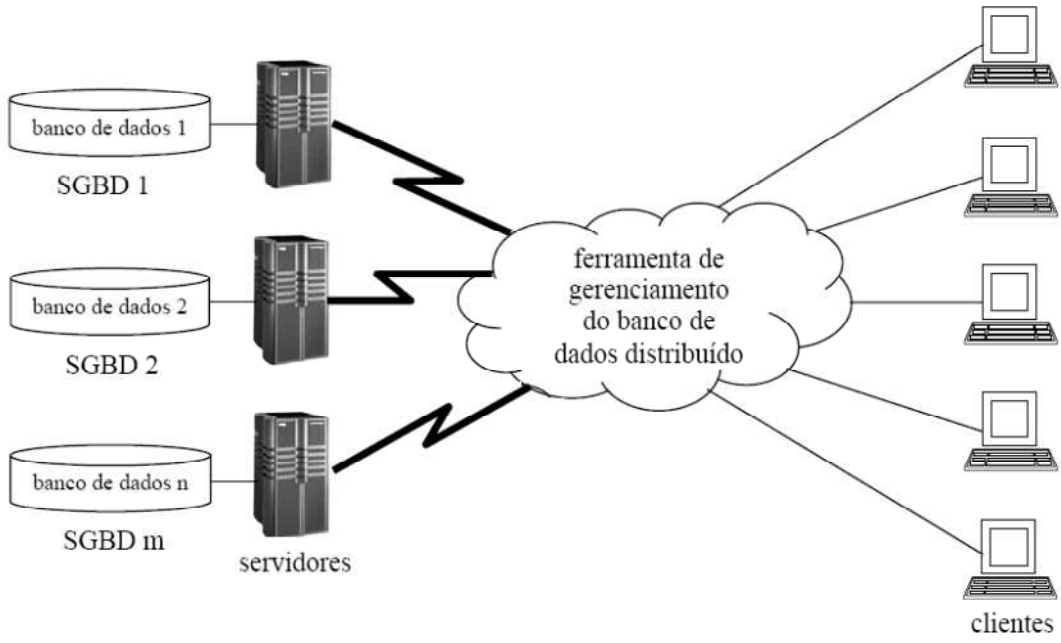

Gráfico 9 - Arquitetura do data warehouse distribuído heterogênio de MOELLER

Cada site nesta arquitetura possui o seu próprio banco de dados autônomo e pode representar um data mart independente. A arquitetura heterogênea possibilita que diferentes SGBD sejam utilizados nos sites da arquitetura, para armazenar os bancos de dados a serem distribuídos. É de responsabilidade da ferramenta de gerenciamento do banco de dados distribuído tratar e oferecer os serviços adicionais voltados ao tratamento da heterogeneidade. Além desses serviços adicionais, as demais funcionalidades da ferramenta de gerenciamento do banco de dados distribuído na arquitetura de data warehousing distribuído heterogêneo são as mesmas funcionalidades oferecidas por essa ferramenta na arquitetura homogênea:

- Conectar os diversos bancos de dados independentes através de uma rede de computadores, oferecendo uma visão lógica integrada dos dados corporativos;
- Atender às consultas dos usuários de SSD que requisitam dados através dos sites da arquitetura; e
- Proporcionar os esquemas de fragmentação e de alocação.

É essencial a presença de um modelo de dados global integrado para o bom funcionamento da ferramenta de gerenciamento do banco de dados distribuído.

5.3.3. Arquitetura de Data Warehousing Distribuído com SGBD Distribuído Único

O Gráfico 10 mostra a arquitetura de data warehousing distribuído proposta por Moeller [MOE01]. Diferentemente do que foi visto nas arquiteturas de data warehousing distribuído homogêneo e heterogêneo, na arquitetura com SGBD

distribuído único não existem banco de dados locais autônomos, ou seja, esta arquitetura não oferece suporte a data marts independentes.

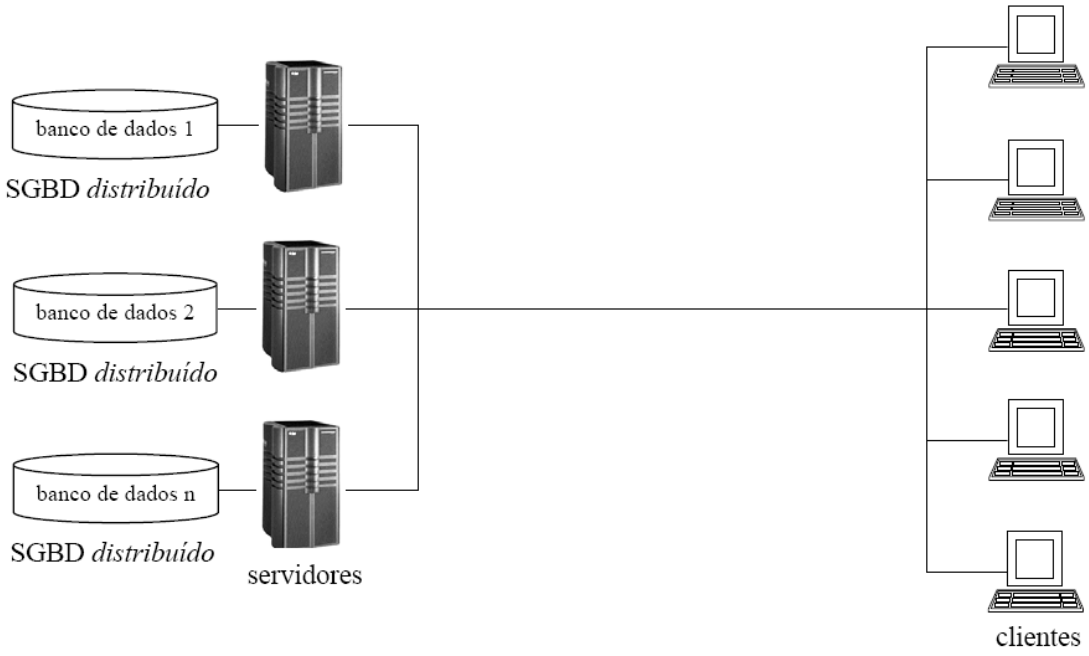

Gráfico 10 - Arquitetura do data warehouse distribuído com SGBD distribuído único de MOELLER

Nesta arquitetura, os dados do data warehouse corporativo podem estar armazenados em diferentes sites, podendo ser distribuídos (fragmentados e/ou replicados) nestes sites à medida que o volume do data warehouse aumenta ou à medida que o número de usuários cresce. O acesso a estes dados é feito através do SGBD distribuído, que desempenha papel similar ao exercido pela ferramenta de gerenciamento do banco de dados distribuído nas arquiteturas de data warehousing distribuído homogêneo e heterogêneo, fazendo-se desnecessária a presença desta ferramenta nesta arquitetura.

Enquanto nas arquiteturas homogênea e heterogênea cada banco de dados local possui o seu próprio modelo de dados individual, na arquitetura com SGBD distribuído único não existem modelos de dados locais. Tal restrição está relacionada ao fato de que

as porções do data warehouse corporativo armazenadas nos diversos sites dessa última arquitetura não podem ser consideradas bancos de dados locais autônomos. Apesar disto, é indispensável a definição de um modelo de dados corporativo na arquitetura com SGBD distribuído único.

6. AVALIAÇÃO GARTNER 2011

Fundada em 1979 por Gideon Gartner, é a líder de pesquisa na área de tecnologia da informação. Realiza avaliações de diversos segmentos de TI anualmente classificando os fornecedores em seus quadrantes quanto à habilidade de executar e abrangência de visão de mercado.

6.1. Critérios

6.1.1. Habilidade para execução

- Produto/Serviço: quão bem o fornecedor atende à gama de funcionalidades de integração de dados exigida pelo mercado, a forma (arquitetura) como essas funcionalidades são entregues e a usabilidade geral das ferramentas. A capacidade funcional do produto é fundamental para o sucesso das ferramentas de integração de dados e, portanto, recebe um peso maior.
- Viabilidade geral. A magnitude dos recursos financeiros do fornecedor e a continuidade de seu atendimento com o cliente e de sua tecnologia. Nesta

iteração do Quadrante Mágico coloca-se uma ponderação alta neste critério para refletir as permanentes preocupações dos compradores sobre os riscos associados com os fornecedores, como resultado das atuais condições econômicas.

- Execução de Vendas / Preços. A eficácia do modelo de preços do fornecedor e de seus canais de vendas diretos e indiretos. Devido aos exames minuciosos sobre as questões de custos e à natureza altamente competitiva deste mercado, aumenta-se o peso deste.
- Receptividade do mercado e histórico. O grau em que o vendedor tem demonstrado a capacidade de responder com sucesso a demanda do mercado por ferramentas de integração de dados durante um período prolongado.
- Execução de Marketing. A eficácia global dos esforços de marketing do fornecedor, o que influencia o grau de "mind share[19]", quota de mercado e fidelidade dos clientes alcançada pelo vendedor.
- Experiência do Cliente. O nível de satisfação manifestado pelos clientes em relação ao suporte de produtos, serviços profissionais e relacionamento geral com o fornecedor, bem como as percepções dos clientes de valor de ferramentas de integração de dados relativos aos custos e expectativas. Nesta iteração do Quadrante Mágico mantém-se o peso elevado deste critério para refletir a forte e contínua preocupação que os compradores estão colocando sobre estas considerações, como resultado das condições econômicas e pressões orçamentais.

Critérios de avaliação	Peso

[19] Termo de marketing. Indica o nível de que certa marca está gravada no subconsciente da pessoa.

Produto/Serviço	Alto
Viabilidade geral	Alto
Execução de Vendas / Preços	Alto
Receptividade do mercado e histórico	Médio
Execução de Marketing	Médio
Experiência do Cliente	Alto

Tabela 3 - Habilidade para execução - Peso dos critérios

6.1.2. Abrangência de visão de mercado

- Entendimento do mercado. O grau em que o fornecedor lidera o mercado em novas direções (tecnologia, produtos, serviços ou outros), e sua capacidade de se adaptar às mudanças de mercado significativas. Dada a natureza dinâmica deste mercado, este item recebe uma ponderação elevada.
- Estratégia de Marketing. O grau em que a abordagem de marketing do vendedor alinha-se com e / ou explora as tendências emergentes e da direção geral do mercado.
- Estratégia de Vendas. O alinhamento do modelo de vendas do fornecedor com a maneira que as abordagens de compra dos clientes preferenciais irá evoluir ao longo do tempo.
- Estratégia de Oferta (Produto). O grau em que mapa do fornecedor estrada produto reflete as tendências de demanda no mercado, preenche lacunas atuais ou fraquezas, e inclui acontecimentos que criam diferenciação competitiva e aumento do valor para os clientes. Além disso, dada a necessidade de ferramentas de integração de dados para suportar diversos ambientes de um

domínio de dados, plataforma e perspectiva mix fornecedor, avaliamos os fornecedores sobre o grau de abertura de sua tecnologia e estratégia de produto. Com o crescimento da diversidade de dados e ambientes envolvidas em iniciativas de integração de dados, este critério recebeu uma ponderação elevada.

- Modelo de Negócios. A abordagem global do fornecedor leva para executar sua estratégia para o mercado de ferramentas de integração de dados.
- Vertical / estratégia da indústria. O nível de ênfase sobre os locais de fornecedores de soluções verticais e profundidade do fornecedor de especialização vertical.
- Inovação. O grau em que o fornecedor tem demonstrado uma disposição para fazer novos investimentos para apoiar a sua estratégia e melhorar as capacidades de seus produtos, o nível de investimento em P & D voltada para desenvolvimento das ferramentas, e na medida em que o fornecedor demonstra energia criativa. Dado o ritmo de expansão das necessidades de integração de dados ea natureza altamente competitiva do mercado, esse critério recebe uma ponderação elevada.
- Estratégia Geographic. A abordagem a presença global que o fornecedor está buscando (por exemplo, presença local direta, revendedores e distribuidores), e estratégia do vendedor e abordagem para expandir seu alcance em mercados além sua região / país.

Critérios de avaliação	Peso
Market Understanding	high
Marketing Strategy	standard
Sales Strategy	standard

Offering (Product) Strategy	high
Business Model	standard
Vertical/Industry Strategy	low
Innovation	high
Geographic Strategy	standard

Tabela 4 - Abrangência de visão - Peso dos critérios

Figura 3 - Quadrantes mágicos de ferramentas de integração de dados Gartner – 2010

6.2. Fornecedores: Pontos fortes e cuidados

6.2.1. Informatica

Figura 4 - Logo Informatica

Redwood City, California, U.S.

www.informatica.com

Produtos: Plataforma Informatica (components incluídos: PowerCenter, PowerExchange, Data Services, Cloud Data Integration)

Base de clientes: mais de 4.200

Pontos fortes:

- Como um dos fornecedores mais amplamente reconhecidos no mercado, Informatica continua a aumentar a sua presença e mantendo sua posição, aparecendo em avaliações de ferramenta de integração de dados mais frequentemente que suas competidoras. A nova versão Informatica 9 está bem alinhada com as demandas do mercado atual e as tendências em evolução. A oferta se expande no fornecimento de associação de dados, e reúne tanto integração quanto a qualidade de dados em uma arquitetura runtime única que se alinha com as tendências de consolidação da tecnologia.
- A grande maioria de clientes estabeleceram Informatica como seu padrão de empresa para ferramentas de integração de dados, e muitos aplicam suas

ferramentas em grande número de projetos que envolvem o uso de um numero de desenvolvedores maiores do que a media. Enquanto quase todos esses clientes aplicam a tecnologia em armazenamento de dados e BI, uma grande porcentagem tem casos de uso adicionais. Especificamente, os dados de migração/transformação e de interfaces entre as aplicações operacionais foram referenciadas por clientes como outras opções de uso da ferramenta. A base de clientes da Informatica continua a expressar um alto grau de satisfação com a relação valor/tempo, o desempenho do suporte do produto, disponibilidade e em geral, no relacionamento com o fornecedor.

Informatica continua como líder no mercado por explorar modelos alternativos de entrega de funções de integração de dados. Embora ainda seja uma pequena parte de seus componentes focados nessas funções alternativas de entrega de dados em relação ao seu modelo padrão baseada em batch, o lançamento do Cloud 9 oferece três tipos de entrega baseadas em serviço. Cloud Services são focados na integração saleforce.com. A plataforma Cloud está disponível para prestadores de serviços independentes e integradores de sistema para desenvolvimento em serviços baseados nas nuvens, e Cloud Edition, uma plataforma da Informatica para implementação em uma nuvem publica definido como Amazon. A mais utilizada das 3 ofertas não padrão parece ser a Cloud Service.

Cuidados:

- Enquanto sua base de clientes reflete seu mix diversificado de casos de uso, a arquitetura de implantação permanece fortemente orientada a batch (entrega

orientada a lotes de dados). A adoção das abordagens nas nuvens e associação de dados é baixa em comparação ao modelo batch.

- O fornecedor enfrenta forte concorrência com a aparição de grandes corporações no cenário (IBM, Microsoft, Oracle e SAP), que oferecem a ferramenta de integração de dados como adicionais a produtos próprios. Informatica precisa crescer seu mind-share e com executivos não pertencentes à área de TI.
- Informatica possui preço alto em relação a muitos competidores. Este precisa continuar a propagar o valor da sua ampla gama de funcionalidades e suas aplicabilidades ou o preço se tornará um inibidor competitivo. Ao oferecer os modelos de entrega baseado nas nuvens e ofertas por demanda ou por prazo, Informatica espera superar esses desafio.

6.2.2. IBM

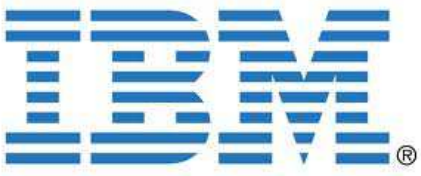

Figura 5 - Logo IBM

Armonk, New York, U.S.

www.ibm.com

Produtos: IBM InfoSphere Information Server (components incluídos: InfoSphere DataStage, InfoSphere QualityStage, InfoSphere Change Data Capture, InfoSphere

Federation Server, InfoSphere Foundation Tools), InfoSphere Data Event Publisher, InfoSphere Replication Server

Base de clientes: mais de 9.000

Pontos fortes

- Os clientes da IBM frequentemente possuem demandas sofisticadas de integração de dados que exigem ferramentas para atender a problemas de integração de dados complexos. IBM continua a demonstrar forte visão do mercado para atender as funcionalidades exigidas, alem de oferecer com sucesso seus componentes de integração a seus clientes existentes. Organizações que adotam as ferramentas InfoSphere tendem a considera-lo como o padrão de integração de dados, refletindo isso, os clientes IBM o utilizam em vários multi-projetos e maior número de desenvolvedores por cliente que a média dos seus competidores. O aumento da base de clientes ocasiona problemas de migração, conversão e operação na interface de dados.
- IBM continua a aumentar o nível de integração e coerência entre seus componentes do InfoSphere Information Server. Lançamentos significativos no final de 2008 e durante 2009 incluem a tecnologia CDC (obtido na aquisição de DataMirror em 2006) com DataStage, as Foundation Tools (para gerenciamento de metadados em vários tipos, perfis e modelos), e a integração do DataStage com InfoSphere MSM Server. A meta para os próximos anos é melhorar a sinergia entre outras tecnologias IBM.
- Distribuição de dados, problemas de sincronização, BI, data warehouse e gerenciamento de dados mestres se beneficiam das funcionalidades do CDC. A comunicação com message brokers, portais Web/Apps e as mais comuns

implantações de banco de dados relacionais são suportados. Enquanto funções similares são oferecidas por alguns concorrentes, a oferta mais competitiva está limitada a necessidades em bulk/batch.

Cuidados:

- Em 2010 houve menor incidência de relatos de problemas em alinhar os diversos componentes, mas ainda há relatos sobre um grande numero de “moving parts” que dificulta a implementação da solução. Como resultado, os clientes relataram a experiência de instalação da solução como “desafiadora”. Apesar destes desafios, a maioria dos clientes indicam que pretendem adquirir novos produtos ou licenças do portfólio InfoSphere nos próximo 12 meses.
- A mesma situação existente em 2009. Durante 2010, a IBM se focou menos em novas funcionalidades e mais em melhorar a qualidade de seus produtos.
- Enquanto a IBM fornece vários pontos de integração entre as tecnologias InfoSphere e o WebSphere de processo e aplicação de integração de capacidades, a maioria dos clientes os utiliza separadamente.
- O preço continua sendo a maior das preocupações para os clientes IBM. O uso da velocidade da CPU como principal parâmetro de preço (adiciona complexidade para os clientes auditarem e modificarem suas implementações) e o relativamente alto custo de uma implementação típica (em comparação com seus concorrentes) criam algumas perspectivas em fornecedores alternativos ou limitar investimentos para um pequeno numero de componentes.

6.2.3. Microsoft

Figura 6 - Logo Microsoft

Redmond, Washington, U.S.

www.microsoft.com

Produtos: SQL Server Integration Services, BizTalk Server

Base de clientes: mais de 10.000

- O principal produto da Microsoft no mercado de ferramenta de dados é o SQL Server Integration Services (SSIS), focada na entrega de dados baseado em batch. Clientes citam o baixo custo total de aquisição, rapidez de uso e capacidade de se integrar com o restante dos recursos do Microsoft SQL Server como principais razões para a escolha de SSIS.
- Clientes reconhecem SSIS como uma ferramenta estável e madura, capaz de suportar em escala empresarial uma implementação um ambiente Microsoft-centric. O grande uso de SSIS dentro da base de clientes SQL Server tem a aumentado a comunidade de suporte, treinamentos e documentações de terceiros sobre as práticas de implementação e abordagens de resolução de problemas.

- O tamanho e presença global da Microsoft fornece uma enorme base de clientes para estudo de melhores práticas, habilidades dominantes e um modelo de distribuição que suporta tanto as vendas diretas quanto a de parceiros. Além disso, clientes relatam um suporte pós-vendas de qualidade, incluindo documentação e mecanismos de apoio on-line.

Cuidados

- Enquanto SSIS pode ser integrado com o BizTalk e a Microsoft pode abordar um estilo de replicação de entrega de dados pelas funcionalidades do SQL Server, essa estratégia não está claramente articulada à visão do mercado.
- A versão do SQL Server 2008 R2 do SSIS ampliou substancialmente a capacidade do fornecedor de suportar vários tipos de requisitos de conectividade de dados. No entanto, a ausência de um sólido CDC (Chance Data Capture) e inabilidade de operar fontes de dados não baseados em SQL Server indicam que o usuário final dessas organizações deve buscar essas funcionalidades em terceiros. Microsoft busca preencher essas lacunas com parceiros.
- Clientes continuam a citar o gerenciamento de metadados (como descoberta de metadados, origem e relatório de dependências) como uma fraqueza substancial. Implementações envolvendo a interoperabilidade entre SSIS e outros produtos (como BizTalk e SQL Server 2008 Master Data Service) são descritos por suas exigências de excessivo esforço em codificação para personalização. Microsoft planeja atender a essas necessidades de integração em uma futura versão do SQL Server 11. Outras lacunas ou deficiências funcionais citados por clientes incluem limitações na qualidade/governança de dados.

6.2.4. Oracle

Figura 7 - Logo Oracle

Redwood Shores, California, U.S.

www.oracle.com

Produtos: Data Integrator, Data Service Integrator, Warehouse Builder, GoldenGate

Base de clientes: mais de 3.500

Pontos Fortes:

- O Oracle Data Integrator (ODI) e os produtos GoldenGate tem como foco central a integração de dados Oracle. A Oracle afirmou que não irá mais dar continuidade ao Oracle Warehouse Builder (OWB). Oracle Data Services Integrator (ODSI) adiciona novas funcionalidades de associação com os produtos Oracle. Alem disso, a aquisição da Oracle GoldenGate Software alcançou seu potencial adicionando classes empresariais e replicação/sincronização à sua suíte.
- Em 2010, clientes relataram a facilidade de uso e a boa curva de aprendizado para o ODI. Os clientes que o utilizam também reportaram a facilidade de integrar da solução com as infraestruturas existentes, aproveitando tanto a

extração, a carga e a transformação (ETL) de dados quanto os sólidos módulos de conhecimento e a boa conectividade.

- A adoção de ambos ODI e GoldenGate continua a crescer dentro do SGBD Oracle e a aplicações baseadas em clientes, mais frequentemente em implementações de ETL tradicional em suporte a BI e data warehouse – no entanto, os clientes Oracle demonstraram recentemente interesse em outros estilos de implementação. A opção por produtos Oracle foi baseada na absorção dos clientes existentes dos produtos GoldenGate anteriores. Clientes citam como escolha dessa ferramenta as funcionalidades para ETL, a boa integração com o SGBD Oracle, a integração com os componentes e as aplicações Oracle Fusion Middleware e a presença geral no mercado e viabilidade.

Cuidados:

- Clientes reportam que para alcançar todas as funcionalidades desejáveis é preciso adquirir vários produtos e com isso os custos sobem. Os preços da Oracle são transparentes nos resultados, mas permanece complexo. Oracle indica que os clientes podem obter um menor preço e licenciamento com o Oracle Data Integrator Enterprise Edition (ODIEE), que inclui tanto o OWB quando o ODI. Custos adicionais são associados ao ODSI e ao GoldenGate e outros componentes adicionais. Finalmente, é importante notar que o ODI 11g foi um esforço conjunto das equipes do OWB e ODI – um lançamento significativo indicando que a integração de dados Oracle está se movendo em direção a uma abordagem unificada de desenvolvimento do produto.
- Uma surpreendente pesquisa encontra relatos de que o suporte da Oracle não parece familiarizado com o ODI e que seus profissionais parecem não ter

conhecimento sobre os produtos da própria Oracle. Surpreendentemente, porque pesquisas da Gartner com cliente relatam a fácil aprendizagem das ferramentas. A combinação com a inquietação dos colaboradores, dos suportes e consultores do produto indica que falta ênfase no treinamento das ferramentas, especificamente do ODI. A Oracle está aumentando seu suporte atualmente em resposta às rápidas adoções do produto no mercado e parece que essa despreparação é um sintoma desse crescimento rápido.

- ODSI é relatado como tendo uma alta incidência de erros de software (e uma falta geral de funcionalidades de depuração), com algumas referencias a componentes mais fracos no conjunto da ferramenta. Isto vai contra os dados da Oracle de apoio ao cliente. É possível que algumas inconsistências estejam ocorrendo entre as métricas de sucesso de suportes e as expectativas dos clientes, o que a Oracle pretende dar uma resposta ao longo do tempo. Clientes ODI reportam um fraco gerenciamento de versão e controles para desenvolvimento (incluindo limitações em gestão de acesso – que são supostamente abordados na versão ODI 11g). Em relação ao grande numero de ferramentas, há uma falta de apoio ao desenvolvimento em equipe. No entanto as vendas das ferramentas de integração de dados Oracle estão crescendo, o que indica uma aceitação do cliente em práticas go-to-market da Oracle.

CONCLUSÃO

Com o crescimento das empresas no mercado atual, a exigência de informação rápida e da maneira mais simples possível está cada vez maior. Isso porque empresas que demoram mais para tomadas de decisões acabam ficando para trás e perdendo competitividade.

Os DW são atualmente a fonte de informação mais importante para as empresas, porém não importa se essa fonte seja de fácil acesso ou se o acesso é rápido se as informações não estiverem lá.

Outro problema que aparece, é que as informações nunca estão em um único lugar, precisando assim integrá-las e convergi-las para um único ponto.

Com isso as soluções de ETL representam uma grande parte do setor de BI das empresas, pois ele é responsável por carregar e padronizar os dados de diversas fontes, e quanto mais rápido for esse processo, mais rápida é a disponibilização da informação para as áreas de gerência para tomadas de decisões e mais rápida será a resposta da empresa às variações do mercado.

Neste trabalho foram abordados os assuntos sobre o histórico de soluções de BI envolvendo DW e os tipos de Banco de Dados (Relacional e Dimensional) que fazem parte das ferramentas que são desenvolvidas para o apoio a tomada de decisão.

Foi apresentado também as formas de distribuição dos DWs para que os mesmos possam atender as demandas de dados que crescem exponencialmente nos dias de hoje. Observamos que um DW unificado pode não suprir as necessidades das empresas com relação ao tempo de resposta, tornando assim as soluções distribuídas cada vez mais presentes.

Por fim foi apresentada a análise da Gartner mostrando o ponto forte e considerações sobre as principais empresas para soluções de BI do mercado na atualidade. Mostrando também que mesmo hoje, com as evoluções tecnológicas ainda existe muito espaço para crescer.

REFERÊCIAS BIBLIOGRAFICAS

ALMEIDA, Maria Sueli e outros, Getting Started with DataWarehouse and Business Intelligence. 1ª edição. San Jose, CA, EUA: International Business Machines Corporation, 1999.

SIMON, Alan R., Data Warehousing for Dummies. 1ª edição. Foster City, CA, EUA: IDG Books Worldwide, 1999.

HAHN, Seungrahn e outros, Capacity Planning for Business Intelligence Applications. 1ª edição. San Jose, CA, EUA: International Business Machines Corporation, 2000.

KIMBALL, Ralph & ROSS Margy, The Data Warehouse Toolkit. 2ª edição. New York, NY, EUA: John Wiley and Sons, 2002.

HAMMERGREN, Thomas C. & SIMON, Alan R.: Data Warehousing for Dummies 2nd Edition. New Jersey: Wiley Publishing, Inc. 2009.

PORTUGAL TELECON, Manual de Introdução Data Warehouse e Informática Power Center 7.

TRABALHOS ACADÊMICOS CONSULTADOS:

“Business Inteligence”

(Trabalho de Conclusão de Curso à Faculdade de Tecnologia São Paulo; Autor: Sandro Hira Pires).

“Estrutura de BI”

(Trabalho de Graduação à Universidade Federal de Pernambuco; Autor: Álvaro Alencar Barbosa Palitot).

WEBGRAFIA

http://www.virtualtechtour.com/assets/GARTNER_DI_MQ_2010_magic_quadrant_for_data_inte_207435.pdf (Acessado em Agosto de 2012).

http://www.kimballgroup.com/ (Acessado em Julho de 2012).

http://pt.scribd.com/doc/86014285/9/Modelo-Relacional (Acessado em Julho de 2012).