

Sistemas de Apoio à Inteligência do Negócio

Asterio K. Tanaka
<http://www.uniriotec.br/~tanaka/SAIN>
tanaka@uniriotec.br



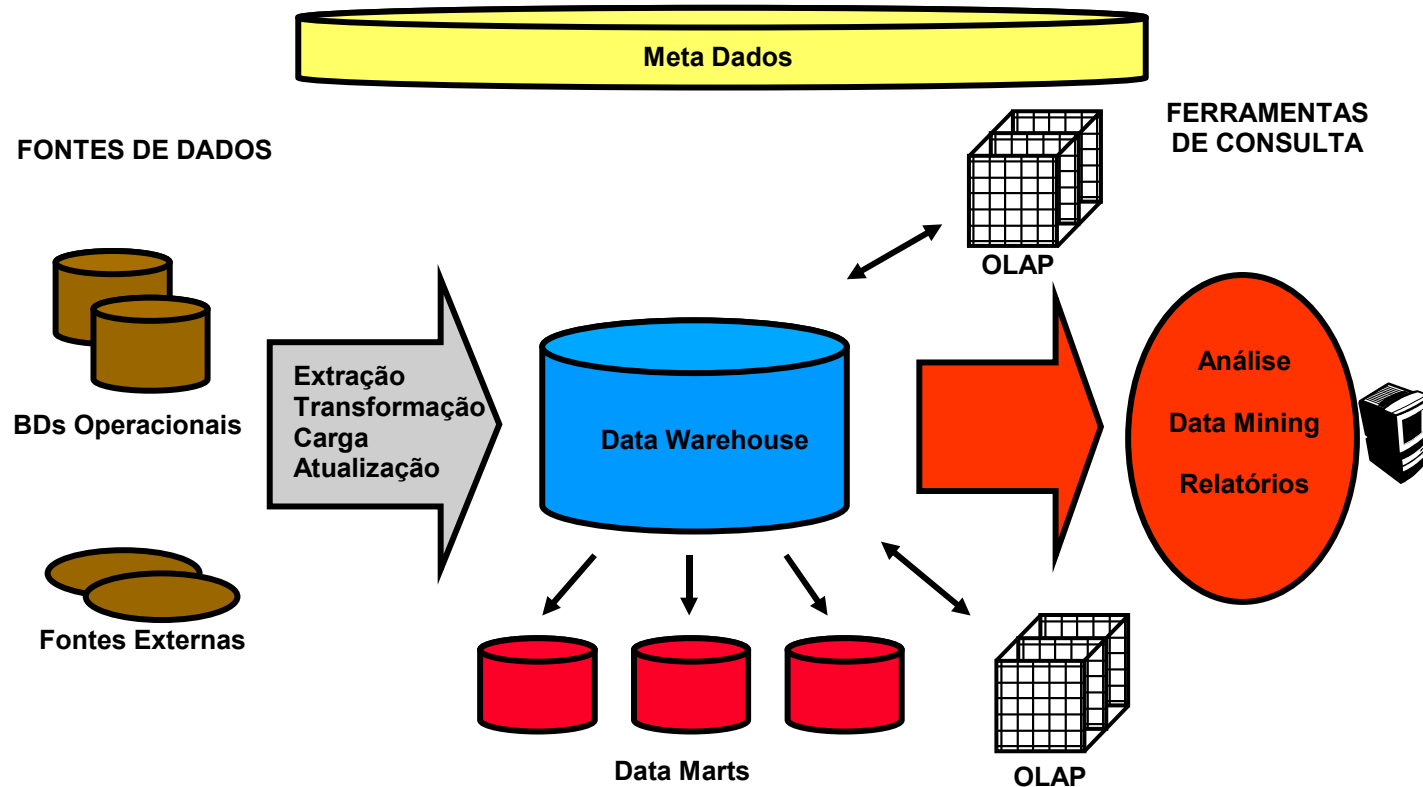
Arquitetura de Data Warehouse

Definição de Data Warehouse

"A Data Warehouse is a
subject-oriented,
integrated,
time-variant,
non-volatile
collection of data in support
of management's decision-making process."

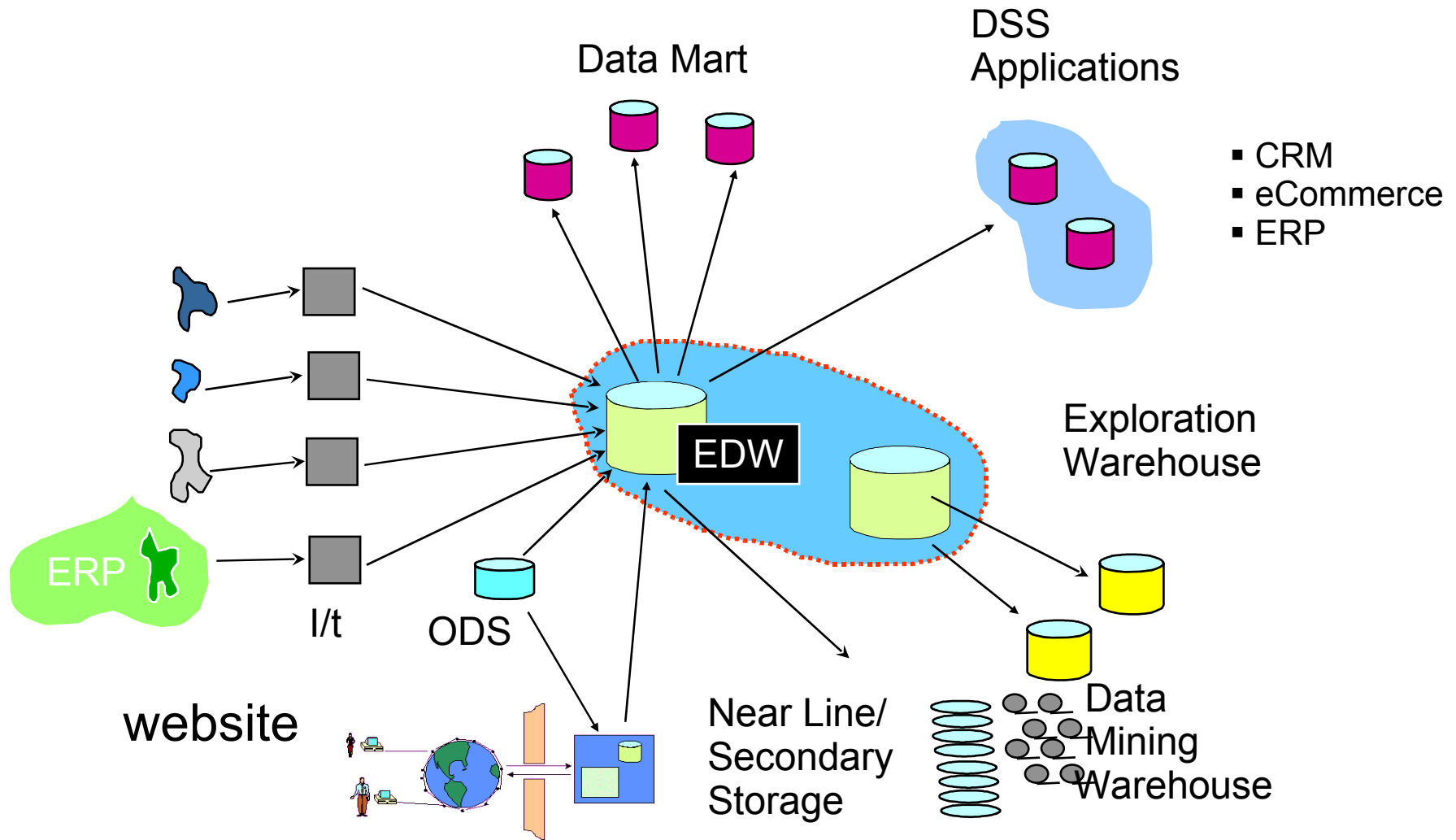
(W. Inmon)

Arquitetura Genérica de um Data Warehouse

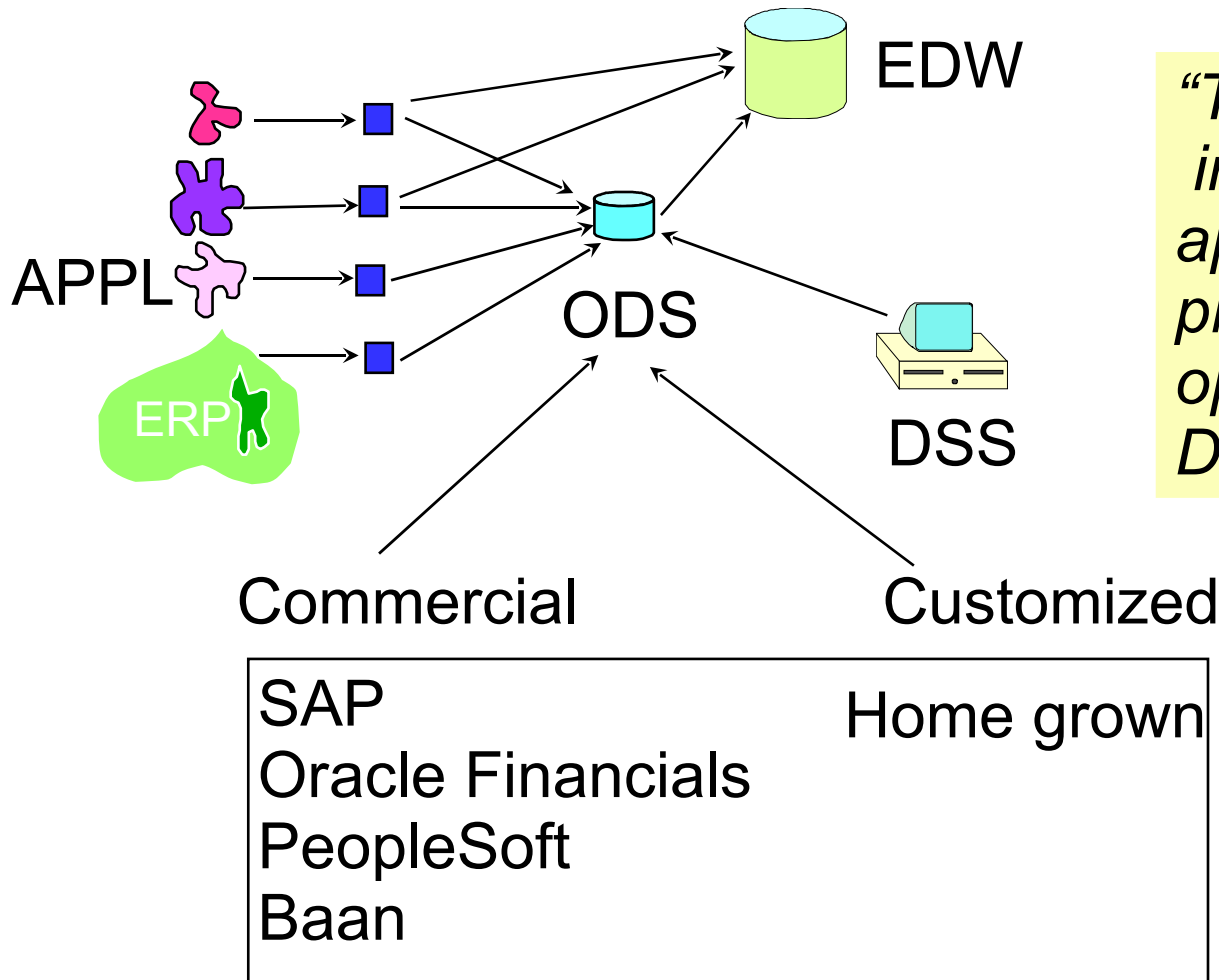


Chaudhri&Dayal, SIGMOD RECORD 1997

Corporate Information Factory de Inmon

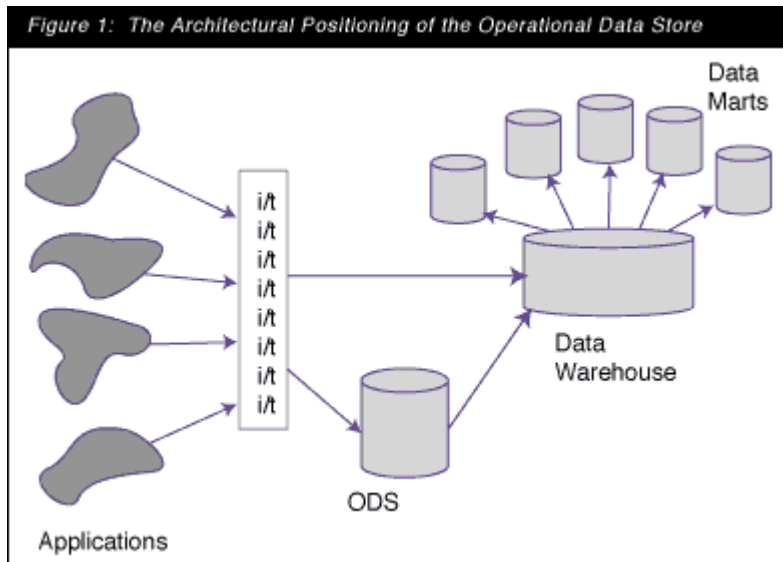


O ODS de Inmon



"The ODS serves to integrate legacy applications and to provide a basis for operational (tactical) DSS processing"

ODS – introduzido por Inmon



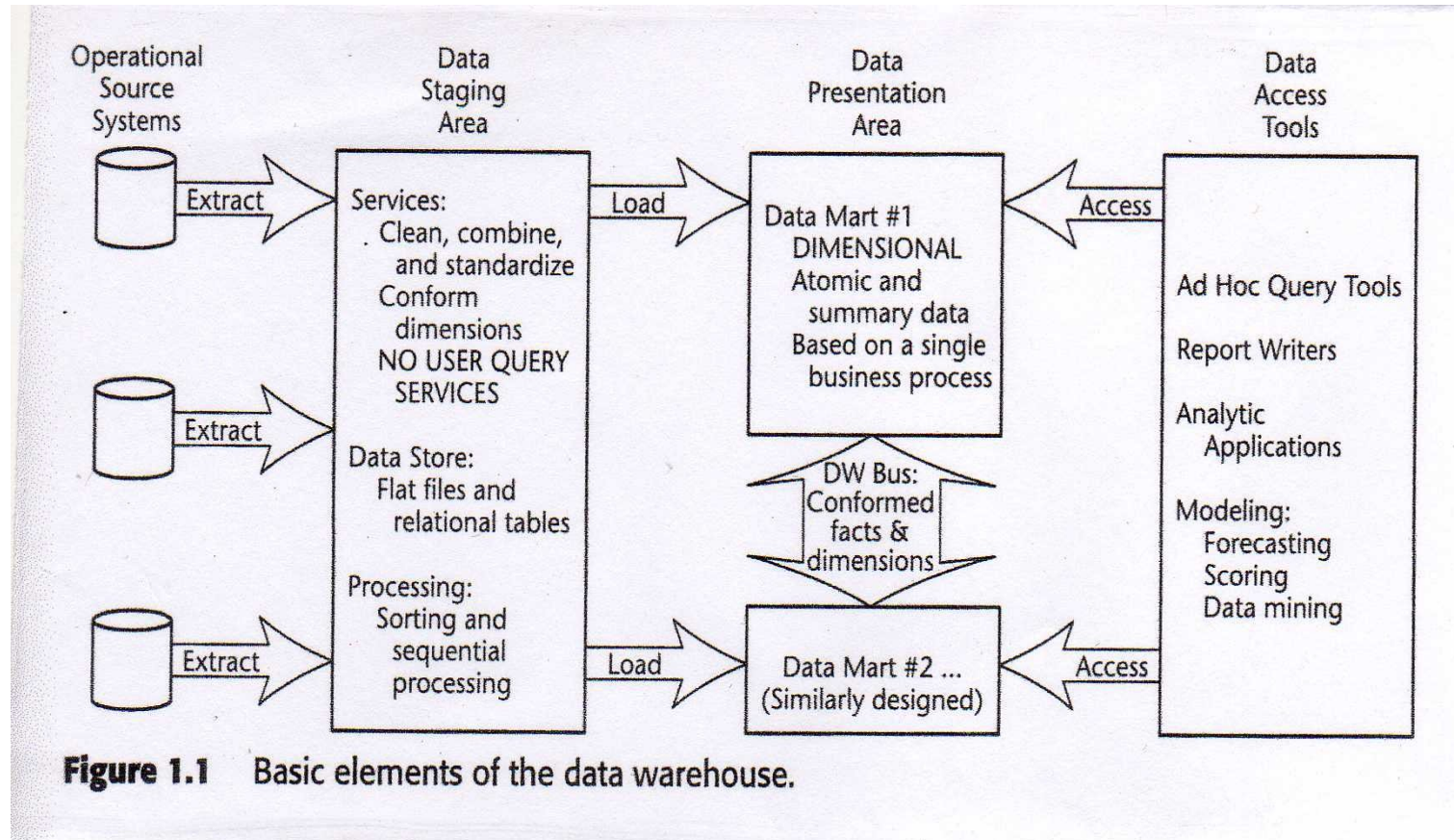
Building the Operational Data Store, W. H. Inmon, Claudia Imhoff and Greg Battas, John Wiley & Sons, 1996
http://www.dmreview.com/article_sub.cfm?articleId=469
(July 1998)

The Operational Data Store (ODS) is a subject-oriented, integrated, current, volatile collection of data used to support the **tactical decision-making** process for the enterprise. It is the central point of data integration for business management, delivering a common view of enterprise data.

The essence of an ODS is the enablement of integrated, collective on-line processing.

- An ODS delivers consistent high transaction performance--two to three seconds.
- An ODS supports on-line update.
- An ODS is integrated across many applications.
- An ODS provides a foundation for collective, up-to- the-second views of the enterprise.
- And, at the same time, the ODS supports decision support processing.

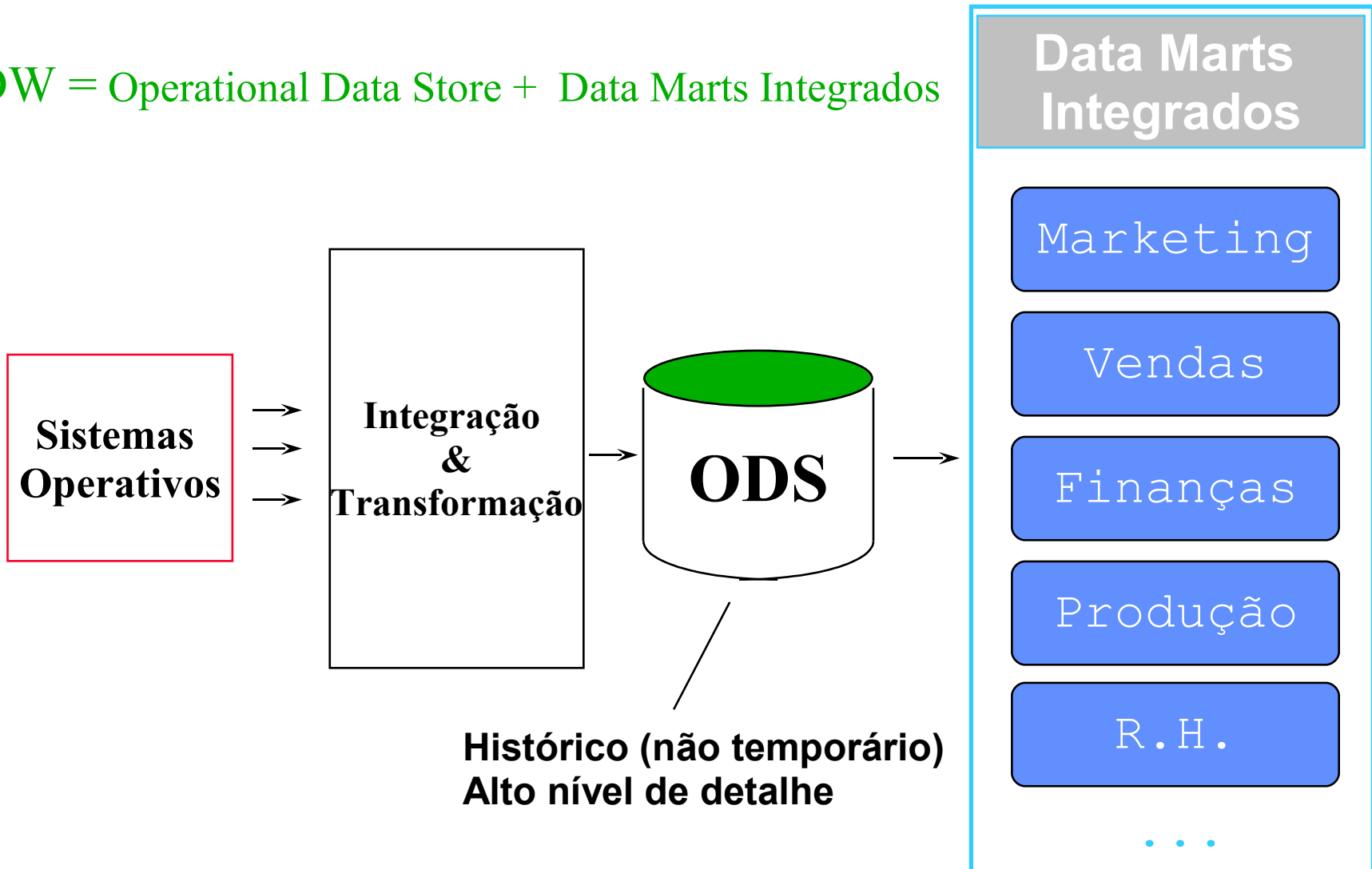
Data Warehouse segundo Kimball



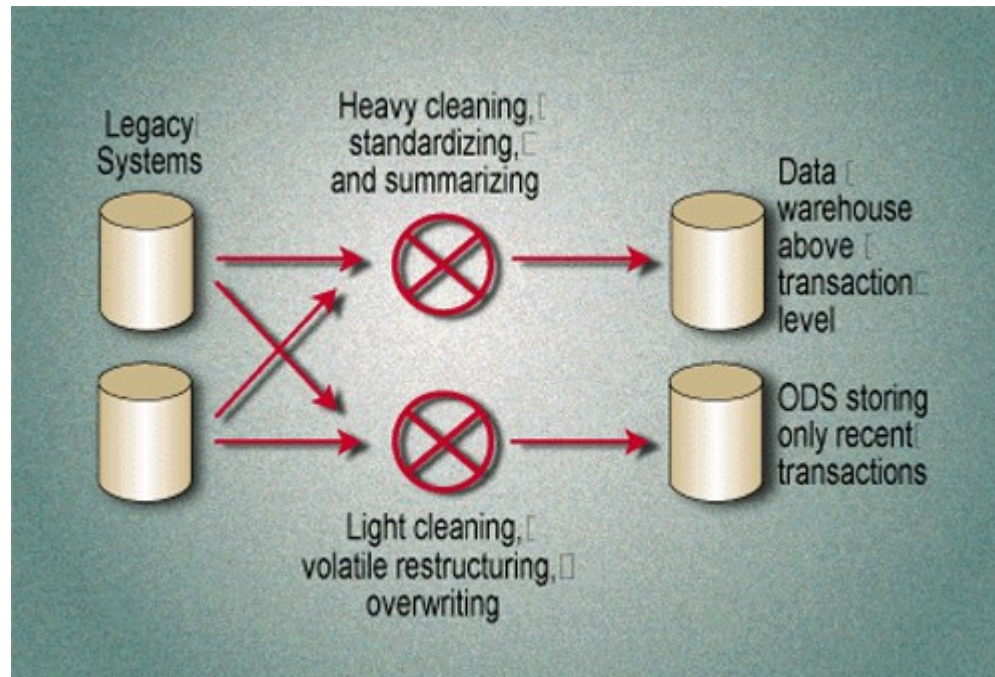
Ralph Kimball, Margy Ross: The Data Warehouse Toolkit, 2ª Edição, Wiley, 2002

Arquitetura de Data Warehouse de Kimball

DW = Operational Data Store + Data Marts Integrados



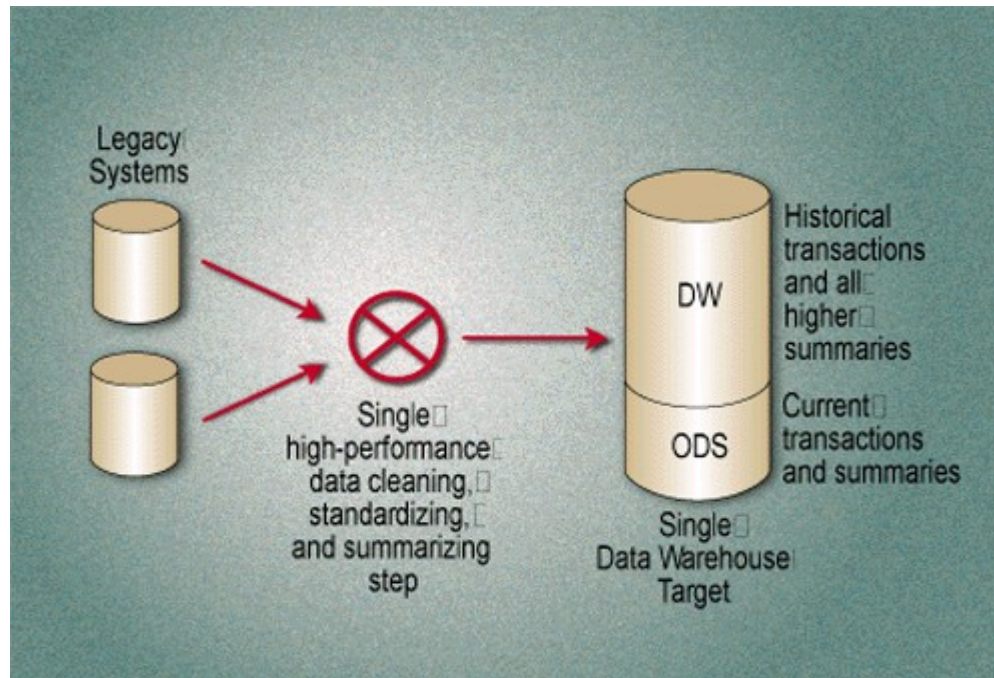
ODS – Inmon by Kimball



The original ODS architecture necessitated two pathways and two systems because the main data warehouse wasn't prepared to store low-level transactions.

<http://www.dbmsmag.com/9712d05.html> (December 1997)

ODS – Kimball



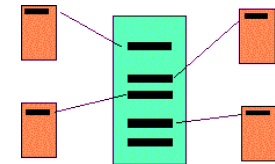
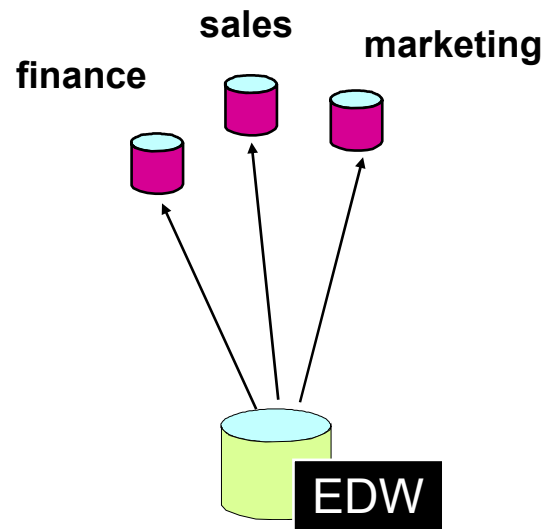
Let us redefine the ODS as follows. The ODS is a subject-oriented, integrated, frequently augmented store of detailed data in the enterprise data warehouse.

The new ODS reality. The cleaning and loading pathway needs only to be a single system because we are now prepared to build our data warehouse on the foundation of individual transactions.

Data Marts de Inmon

Data Mart

- OLAP
- Multidimensional



Estruturas
Star ou Snowflake

Para Inmon, Data Marts são depósitos secundários derivados do Data Warehouse.

Onde Inmon e Kimball concordam

- **É necessária uma arquitetura**
- **Dimensões compartilhadas e medidas definidas através de todas as áreas da empresa**
- **O esquema estrela é útil para apresentar informações aos usuários**
- **Construir o DW iterativamente (mesmo que projetado globalmente)**
- **Metadado é fundamental**
- **Cada um acredita(va) estar certo e o outro errado!**

Onde eles discordam (ou discordavam)

- **Alguns pontos da arquitetura**
- **Qual modelagem usar e onde**
 - ER / Relacional (normalizado)
 - Star Schema/Modelagem Dimensional
- **O papel dos Data Marts**
- **Abordagem de projeto e construção do DW**

Componentes da Arquitetura

Bill Inmon

- Camada de Transformação/Integração
- Data warehouse corporativo
- ODS corrente
- Data Mart
- Exploration DW
- Metadado

Ralph Kimball

- Data staging area
- Coleção de DM's = DW
- ODS histórico
- Data Mart – papel diferente
- Dados arquivados
- Metadado

DW e o papel da modelagem E/R

Bill Inmon afirma

- ER Model is suitable for data warehouses because it is stable, and supports consistency and flexibility
- Normalised data is ideal basis for the design of the Data Warehouse and the ODS
- May not be suitable for the data mart, which deals heavily with regular query activity and time-variant analysis

Ralph Kimball afirma:

- ER Models are too complicated for end users to understand
- ER Modeling/normalising only suitable for OLTP or in data staging area since it eliminates redundancy
- Results in too many tables to be easy to query
- ER models are optimised for update activity not high performance querying

Modelagem Dimensional e Star Schema

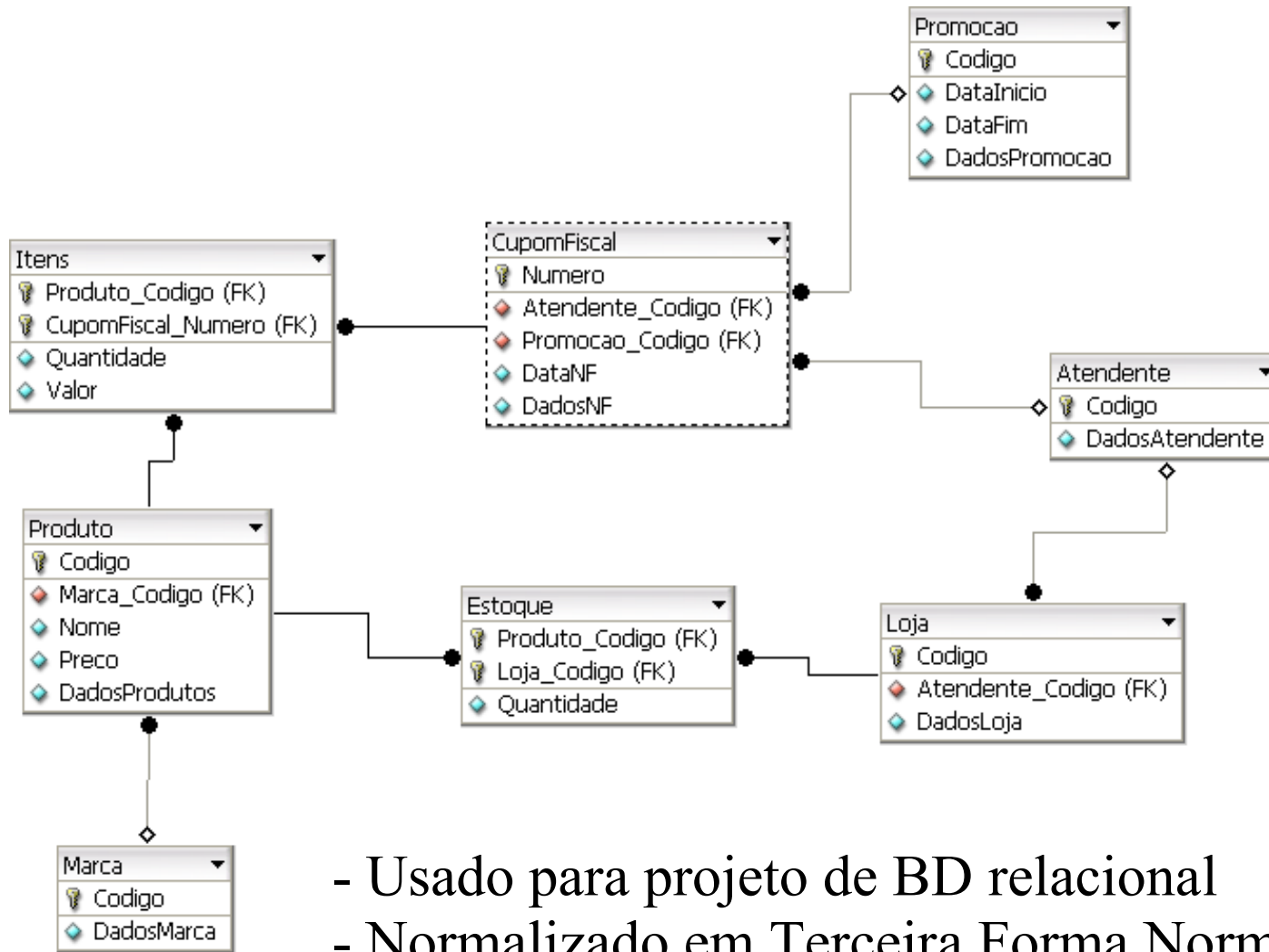
Bill Inmon afirma:

- Dimensional Modeling is reasonable viable technique for designing data marts, when type of access is very predictable
- Dimensional models are not suitable for updating at all
- Differing business areas will likely want a different dimensional model to look at similar data
- Series of dimensional models are not flexible enough to support an enterprise's entire Data Warehouse

Ralph Kimball afirma:

- Dimensional Modeling is the only viable technique for designing databases in the Data Warehouse environment because it provides a predictable framework
- Even lowest level granular data should be in dimensional format
- Every ER model has an equivalent dimensional model representation
- Any type of business data can be represented as a “cube”

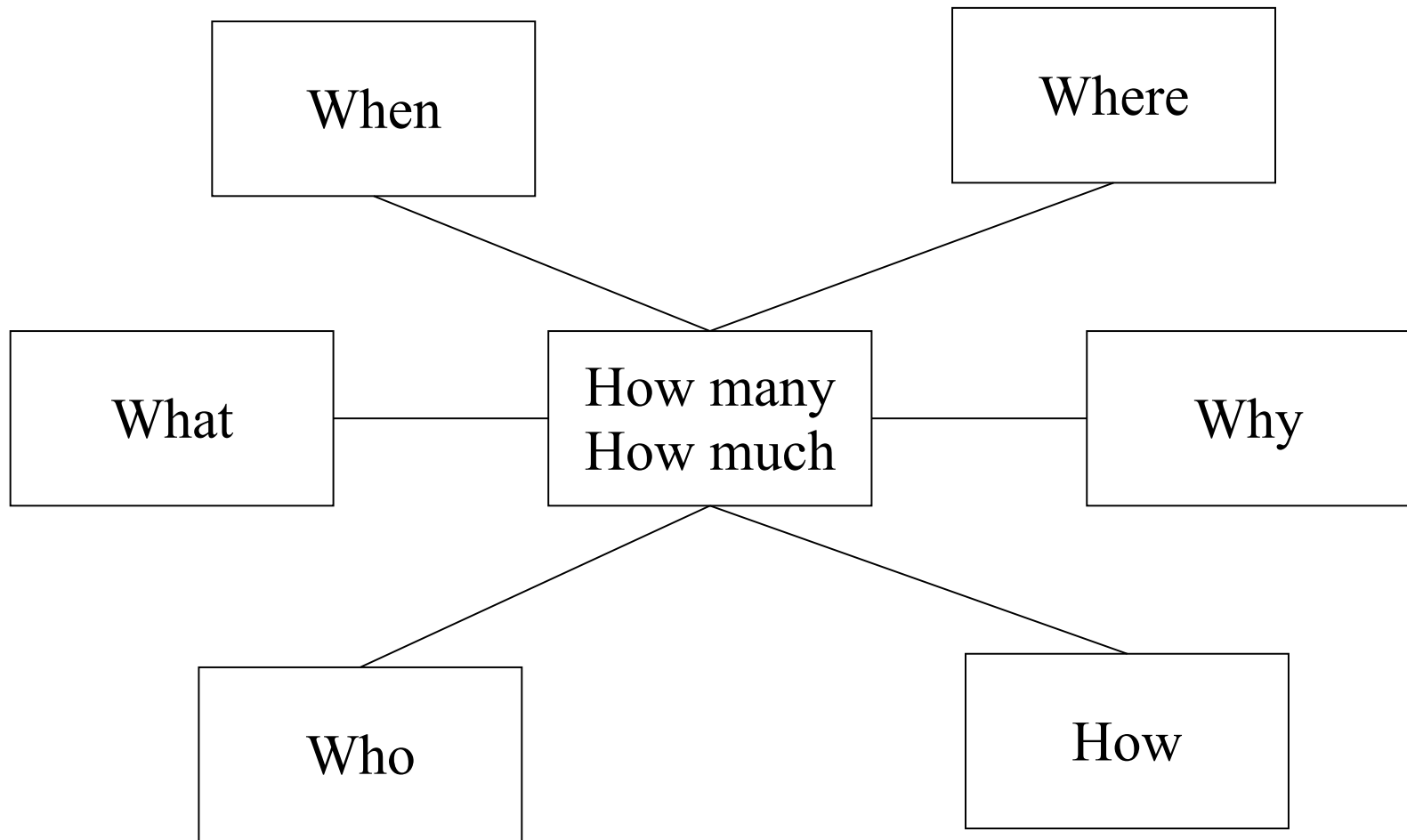
Modelo Entidades Relacionamentos



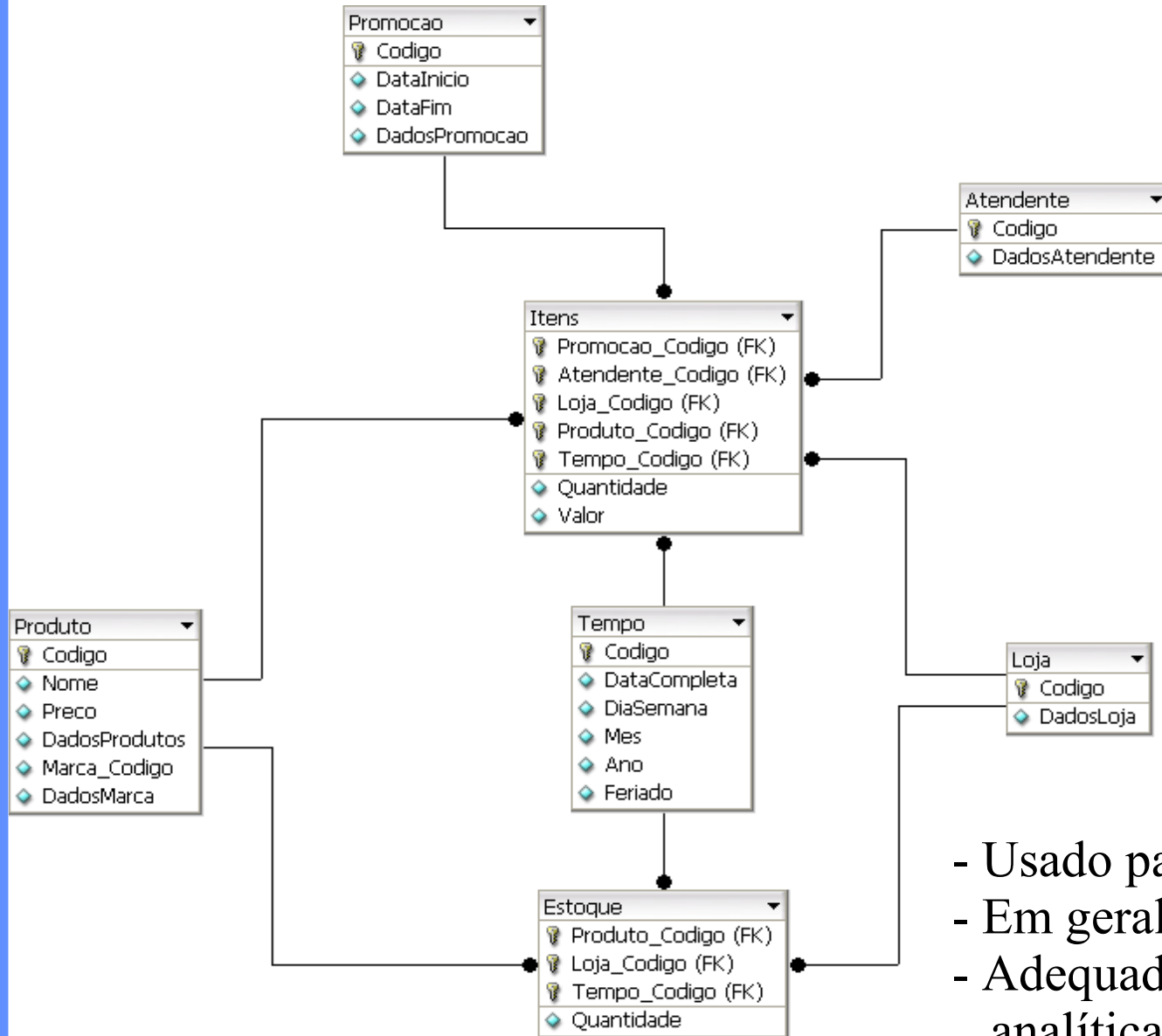
- Usado para projeto de BD relacional
- Normalizado em Terceira Forma Normal
- Adequado para aplicações operacionais (OLTP)

Esquema Estrela de DW

5 W e 3 H



Modelo Dimensional (Esquema Estrela)



- Usado para projeto de DW/DM
- Em geral desnormalizado
- Adequado para aplicações analíticas (OLAP)

Comparação entre Modelos

• Modelo ER/Relacional

- Mais complexo
- Anos 70 – BD relacional
- Tabelas representam conjuntos de entidades e relacionamentos
- Tabelas resultantes naturalmente normalizadas (até 3a Forma Normal)
- Tabelas acessadas indistintamente de filtro inicial
- Maior necessidade e dificuldade de junção
- Maior dificuldade de leitura e de consulta por usuário não especializado

• Modelo Dimensional

- Estrutura mais fácil e intuitiva
- Anterior ao ER, recriada por Kimball
- Tabelas representam Fatos e Dimensões
- Tabelas Fato normalizadas, Tabelas Dimensão podem não ser normalizadas
- Tabelas Dimensão são pontos de entrada para acesso
- Junções só ocorrem entre Tabelas Fato e Dimensões
- Leitura e consulta mais fáceis para usuários não especializados

Papel do Data Mart

Bill Inmon afirma:

- “Data marts should be populated by the data warehouse and external data only
- Can contain subsets, aggregated data or atomic data
- Provide a departmental view of the world
- May or may not reside on a different platform from DW
- Provide for repeatable, predictable types of information delivery”

Ralph Kimball afirma:

- “Successive data marts built on a 'star schema model' together form a data warehouse
- The bad publicity about data marts comes from implementation of isolated stovepipe data marts done badly, and not conforming dimensions and measures
- Data Marts can be atomic but should still be in dimensional view format”

Abordagens de projeto de DW

- **Inmon: Top Down**
 - Data Warehouses Corporativos
 - » de grande abrangência
 - » complexos
 - » alta probabilidade de insucesso
- **Kimball: Bottom Up ou Middle Out**
 - Data Marts Setoriais
 - » Marketing, Financeiro, Administrativo, etc.
 - » Projetos evolutivos
 - » Enfoque inicial nos aspectos mais críticos
 - » Aproveitamento da estrutura operacional disponível
 - » Retorno mais rápido
 - » Acúmulo de experiência : menor risco e menor custo

Onde Inmon “funciona”

- **Grandes organizações com muitas unidades de negócio diferentes que precisam compartilhar informações**
- **Multiplos SSD's utilizados, e inconsistência entre eles é sentida.**
- **Modelagem tradicional é uma prática e é bem compreendida**

Onde Inmon “falha”

- **Pouca atenção a detalhes de modelagem**
- **Não enfatiza importância de dimensões compartilhadas e medidas uniformes**

Onde Kimball “dá certo”

- **Pequenas organizações , capacidade de medida previsível**
- **Lugares mais estáveis**
 - **Dimensões e medidas são bem conhecidas e não mudam com frequência**
 - **Onde grão pequeno não gera Terabytes**

Onde o Kimball “falha”

- **Se escolher a granularidade errada da primeira vez ...**
- **Se surgir uma nova maneira de olhar o negócio, pode custar um outro projeto**
- **Assume que usuários não conseguem lidar com um snow-flake (Terceira forma normal)**

Convergência de Abordagens

“Why not....

- Pay strict attention to conforming dimensions and measures across the business**
- Also model hierarchies early in piece**
- Have a permanent staging area (3rd normal form) and name it an atomic data warehouse**
- Feed dimensional data marts from this DW/Staging area**
- Build data marts for departments going through staging area”**

Abordagem corrente

(mix de Inmon e Kimball, mais para Kimball)

- **Estratégia**
 - Desenvolver incrementalmente
 - Visão Integrada
 - Dividir para conquistar
 - Errar pequeno
- **Implementação**
 - Planejamento Top-Down
 - Desenvolvimento Bottom-Up (ou Inside Out), um Data Mart de cada vez, resultados devem ser atingidos em pequenos ciclos (ex.: a cada 3 meses)
 - Cada Data Mart deve ser encarado de forma evolutiva
- **Desafio**
 - Garantir a coerência entre os vários Data Marts, através de dimensões conformadas.

Fatores Críticos de Sucesso em Projetos de DW/DM

- Foco bem definido
- Patrocinador forte
- Existência dos dados necessários
- Envolvimento dos usuários
- Qualificação da equipe de projeto
- Arquitetura tecnológica bem definida
- Marketing interno e acompanhamento
- Gerência e manutenção de metadados

Componentes Potenciais do Ambiente de DW

1. Repositório de Metadados
2. Ferramentas de Projeto CASE
3. Ferramentas de Extração, Transformação e Carga (ETL)
4. Ferramentas para Qualidade e Limpeza
5. Ferramentas para Replicação
6. Provedores de Interfaces de BD ODBC/OLE
7. Ferramentas de Gateway para BD Legados
8. Bancos de Dados Relacionais
9. (Bancos de Dados Não-Relacionais Legados)
10. Bancos de Dados Multidimensionais
11. Ferramentas OLAP
12. Ferramentas de Relatório e Consulta
13. Ferramentas de Data Mining
14. Ferramentas de Monitoramento e Controle
15. Pacotes de Aplicação para Data Warehouse

**Todos estes
componentes
manipulam/geram
metadados**

Transporte de Dados (Data Staging)

- **Extração**
 - Coleta de dados nos sistemas existentes
 - Operação demorada e complexa
 - Muitas vezes, desenvolvimento ad-hoc (caso a caso)
- **Transformação e Limpeza**
 - Transformações para clareza e integração
 - » Recodificação de categorias: (m/f, masculino/feminino para M/F)
 - » Alterações e uniformização de unidades de medida, nomes de campos, formatos de datas...
 - » Escolha de chaves das tabelas (Surrogate Keys)
 - Limpeza para qualidade da informação extraída
 - » Correções de ortografia, conflitos de domínios, tratamento de valores ausentes
 - » Combinação de dados de fontes múltiplas, eliminação de duplicatas
- **Carga e Realimentação**
 - Trade-off: muito frequente é “caro”, pouco frequente significa dados “velhos”

Ferramentas de ETL

- Deve-se considerar “desenvolver versus comprar”:
inicialmente, muitas empresas escrevendo seus próprios programas;
- Produtos incluem geradores de código ou “transformadores proprietários”;
- Muitas ferramentas são voltadas para áreas específicas, embora com funcionalidades em comum;
- A maioria dos produtos é relativamente imatura, embora tenham melhorado muito no último ano;
- As ferramentas são geralmente muito caras, embora um novo modelo de preços esteja surgindo;
- Estas ferramentas são mais adequadas para ambientes complexos (múltiplas fontes e destinos, muitas transformações, muita limpeza necessárias), desde que as transformações não sejam muito complicadas.

Dados, Metadados, Meta-metadados

- **Dado** é uma descrição de alguma “coisa”.
- **Metadado** é um tipo de dado em que a “coisa” sendo descrita é um dado.
- **“Meta-”**
 - significa “algo que descreve ...”
- **“Meta-Meta-”**
 - significa “algo que descreve algo que descreve...”
- **Metadados**
 - “dados sobre os dados”
 - quaisquer informações que permitam identificar, localizar, utilizar e entender os dados
 - "Metadata is structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities." [American Library Association, Task Force on Metadata Summary Report., June 1999]

Vide Meta Meta Data Data - Making a List of Data About Metadata and Exploring Information Cataloging Tools. Ralph Kimball, 1998

<http://www.fortunecity.com/skyscraper/oracle/699/orahtml/dbmsmag/9803d05.html>

Dado e Metadado: onde está a fronteira?

Um exemplo para reflexão:

Valor de ações

Companhia	Ano	Valor
XX	1995	20
XX	1996	30
YY	1995	10
....		

Valor de ações

Companhia	1995	1996	1997	1998
XX	20	30	26	40
YY	10	15	8	23
....				

Valor de ações 1995

Companhia	Valor
XX	20
YY	10
....	

Valor de ações 1996

Companhia	Valor
XX	30
YY	15
....	

...

Gerência de Metadados

- **Grande desafio na construção e manutenção de um DW**
 - Formatos de dados inconsistentes
 - Dados inexistentes ou inválidos
 - Diferentes níveis de agregação
 - Inconsistências semânticas
 - Qualidade de dados e janela de tempo
 - Acesso global (distribuído e replicado)
 - Administração e controle
- **Integração do DW com outras ferramentas aumenta o problema**

Diferentes tipos de metadados

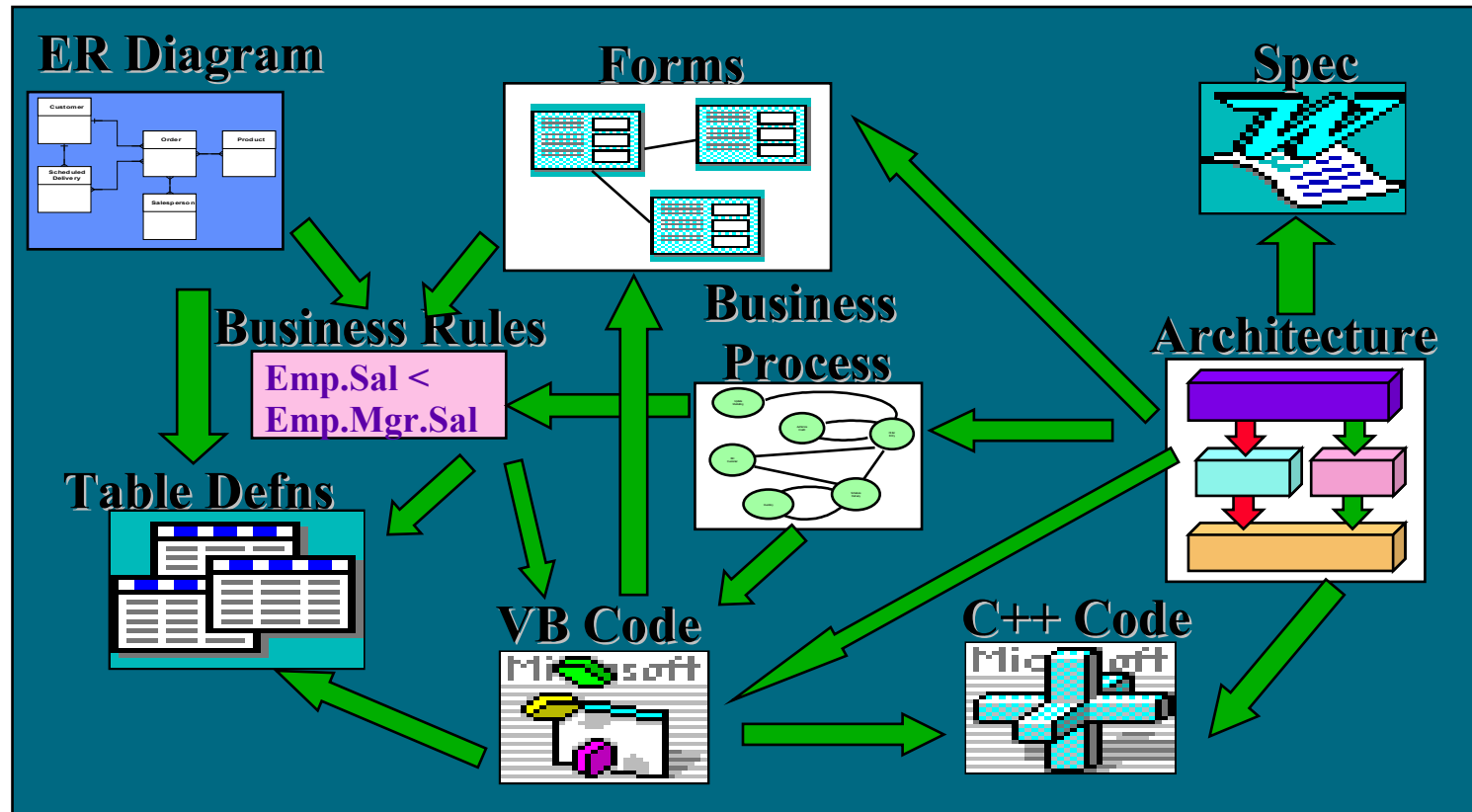
- **Metadado Técnico e Administrativo**
 - altamente estruturado
 - informações com definições, transformações, gerência e operação
 - geralmente tratável via uma ferramenta de repositório
- **Metadado de Negócio**
 - tanto não-estruturado quanto estruturado
 - mais difícil de ser tratado e integrado por uma ferramenta altamente estruturada tipo um repositório
- **Metadados em BI**
 - Metadados para ETL – Extração, Transformação e Carga de sistemas OLTP
 - Metadados de OLAP: descrições de cubos, dimensões, medidas, hierarquias, níveis
 - Metadados de ferramentas front end: rótulos de telas e relatórios
 - Metadados de Data Mining: descrições de algoritmos, consultas, resultados

Importância de um Repositório

- **Repositório**
 - » ferramentas que provêem armazenamento e funcionalidade de gerência e acesso a metadados
- **Visão global e integrada de metadados**
- **Gerenciamento do ciclo de vida dos metadados**
- **Integração com ferramentas de outros fornecedores**

Repositório = Depósito Genérico de Metadados

- Um BD de informações sobre artefatos criados, global através de ferramentas.



Metamodelos

- **Repositórios são desenvolvidos sobre Metamodelos, isto é, metadados de modelos.**
- **Um modelo deve estar conforme com um metamodelo.**
- **Camadas de metamodelos, segundo o OMG (Object Management Group):**
 - **M0: objeto instância, linha de tabela, registro (ex: “João da Silva e seus dados”)**
 - **M1: modelo, esquema (ex: classe UML ou tabela de banco de dados “Cliente”)**
 - **M2: metamodelo (ex: Unified Modeling Language - UML, Common Warehouse Metamodel - CWM, Knowledge Discovery Metamodel – KDM)**
 - **M3: meta-metamodelo (ex: Meta-Object Facility - MOF)**