Data Warehouse

1º Semestre

Mestrado em Desenvolvimento de Software e Sistemas Interactivos

Instituto Politécnico de C.Branco

11ª Edição – 2021/22

Slides #2 - Componentes de um DW

Eurico Lopes

V.09-27.09.19

1

Indice

- **Dados vs. Informação**
  - Dispersão das Fontes de Dados
  - Necessidade Convergência + Navegação
- **Separação Operações / Analítica**
  - OLTP vs.. Data Warehousing
  - Modelação E&R não é a Chave de tudo
  - Modelagem do Tempo
  - DW como "Peça" de Middleware
  - DW vs. Meta-Arquitetura Organizacional
- **Aplicações vs.. Business Intelligence**
  - Soma de Componentes ou Conceito
  - Níveis da Arquitetura
  - Fontes Operacionais
  - Área de Retenção (*Data Staging Area*)
  - *Operacional Data Store* (ODS)
- **Modelagem Dimensional**
  - Data Mart
  - Arquiteturas Alternativas de DW
  - OLAP: MOLAP, ROLAP e HOLAP
  - Classes de DSS
- **Projeto Data Warehouse**
  - Fatores Críticos de Sucesso
  - Ciclo de Desenvolvimento
  - Gerir um Data Warehouse
  - DW Manager
- **DW Gurus e suas Metodologias**

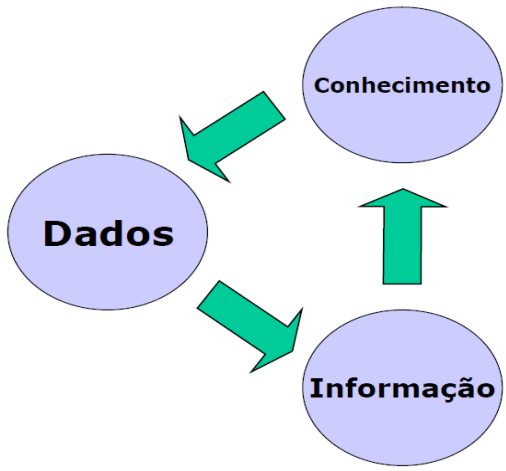
2

Componentes de um DW

- **Objetivos**
  - **Depois de concluir este tópico, você deve ser capaz:**
    - Identificar aa fontes de dados que alimentam um DW;
    - Identificar o conceito de *Operações* numa organização e o processo de Decisão Analítica;
    - Perceber o conceito de *Business Inteligence* (BI)
    - Identificar os componentes que constituem um DW;
    - Perceber o significado dos termos: OLAP: MOLAP, ROLAP e HOLAP;
    - Identificar os fatores sucesso num Projeto de DW;
    - Perceber as diferenças entre as abordagens de Kimball e Inmon.

3

DW – Dados vs. Informação



- Características da Informação
  - Relevância
  - *Timing*
  - Precisão
  - Orientação à ação

4

DW – Pronto a Fazer vs.. Pronto a Usar

Jumbo	**Pans & Co**
(Normalização)	(*Desnormalização*)
Ingredientes base *Condimentos* *Comprar* *Fazer* *Comer*	**Sandwiches** *Comprar* *Comer*

Otimizado para Operações

Optimizador para Consumidores

5

DW - Dispersão das Fontes de Dados

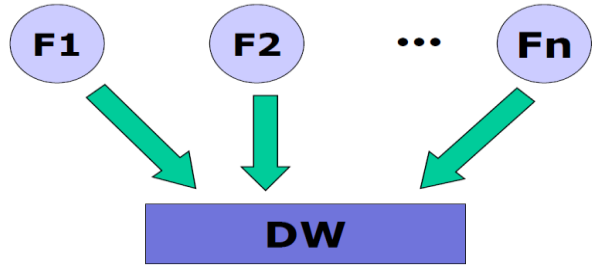
BANCA	**RETALHO**	**TMN**
Back Office Front Office Call Center Gestão de Produtos (especifícos)	Front Office ERP Entreposto Finantials	Billing Customer Care Engenharia (rede) Finantials

Sistemas Heterogéneos
Multiplicidade de Aplicações
Diversidade de Interfaces

Orientação Operacional
Visibilidade Dificultada

6

DW – Necessidade Convergência + Navegação



Convergência é uma necessidade imperativa

Um sistema, Uma Aplicação, Um interface

Visibilidade = Convergência + Navegabilidade

Data Browsability

7

DW - Separação Operações / Analítica (1)

Ilhas de dados
Foco nos dados atuais
Estruturas de dados complexas
Carga transacional (OLTP)

Dados consolidados
Foco nos dados históricos
Estruturas de dados intuitivas
Carga de interrogação

8

DW - Separação Operações / Analítica (2)

Necessário isolar o impacto das
explorações analíticas das operações

Nasce a "Janela Noturna" do Batch de Replicação !

Este especto, embora tático, permanece como
fundamental para justificar o investimento num DW

9

DW – OLTP vs. Data Warehousing

	OLTP	DW
Objectivo	Operar negócio	Analisar negócio
Interacção	Pré-definida	*Ad hoc*
Tipo iteracção	Transacção	*Query*
Operações	*Read/Write*	*Read* (99%)
Registos manipulados	Dezenas	Milhões
Tipo acessos	Indexação	*Table scans*
Conteúdo	Dados atómicos	Dados consolidados, calculados, sumarizados
Tamanho	M-Gbytes	G-Tbytes
Tipo Utilizadores	Mecânicos	Especialistas
N.º utilizadores	x100-x1000	x10-x100

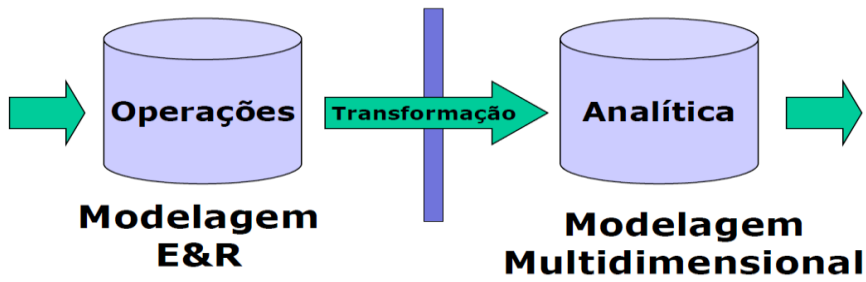
10

DW - Modelação E&R não é a Chave de tudo

Nas Operações a 3ª forma normal é Lei !

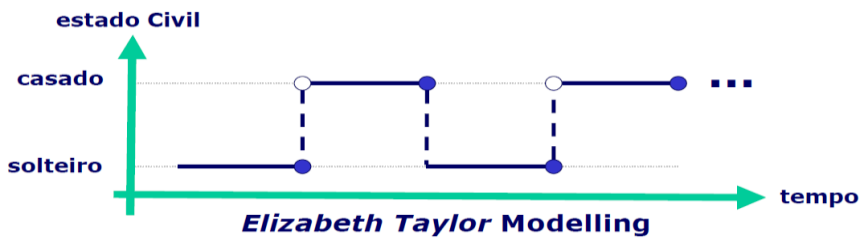
Os puristas já vão na n-ésima forma normal ...

E na área analítica ?



11

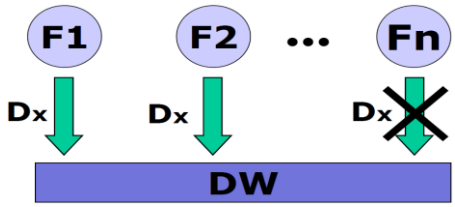
DW - Modelagem do Tempo



- Transições de estado não são, normalmente, mantidas nas bases de dados operacionais. Estas só guardam o último estado.
- No entanto para propósitos analíticos essas mudanças não são desprezáveis !
- O tempo é uma dimensão fundamental num DW ...

12

DW - Resolução de Incoerência de Dados

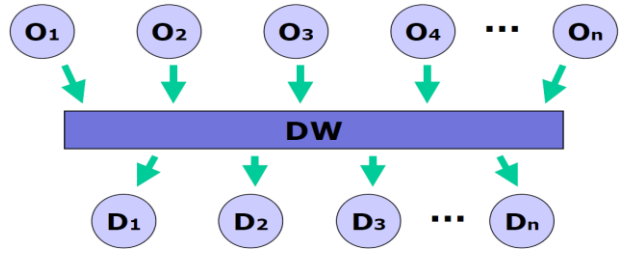


- **Situações de informação operacional incompleta (Dx em F1 e F2),**
- **ou potencialmente incorreta (Dx em Fn).**

- O DW só pode ter uma versão da verdade! **Como resolver o problema**?
  - Definições claras e horizontalmente partilhadas
  - Validação semântica dos conteúdos face às definições
  - Seleção ou combinação das fontes eleitas.
- **DW = Oportunidade de definição de Metadata**
  - (Dados sobre os Dados) Organizacional

13

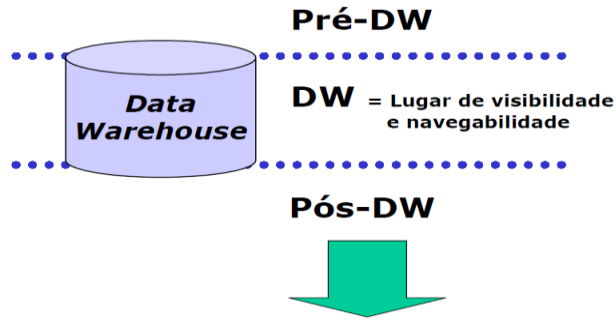
DW - O DW como "Peça" de Middleware



- O DW é invisível para o utilizador final !
- O DW é uma componente de intermediação – um ***Broker*** de dados.
- Neste sentido é possível denominá-lo de ***Middleware***.

14

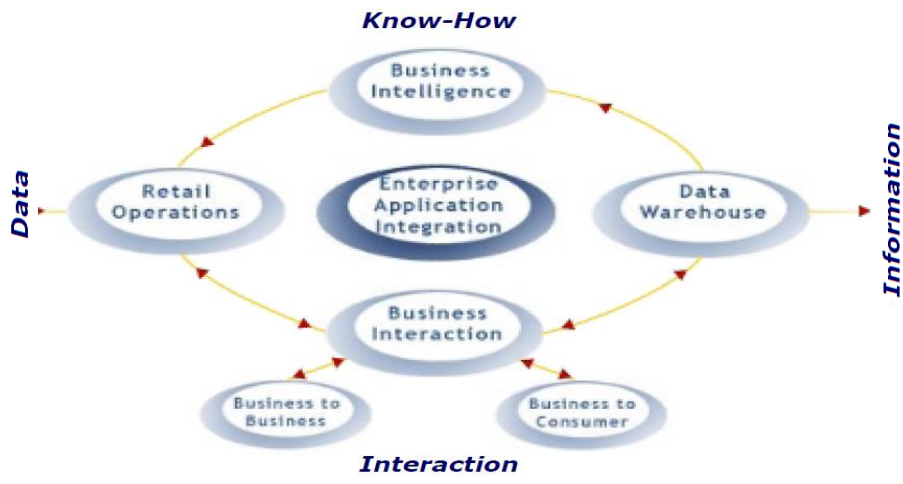
DW - DW vs.. Meta-Arquitetura Organizacional



- O DW surge como *enabler* de um novo conjunto de aplicações até agora impossíveis de construir.
- Soluções inteligentes passivas ou ativas!

15

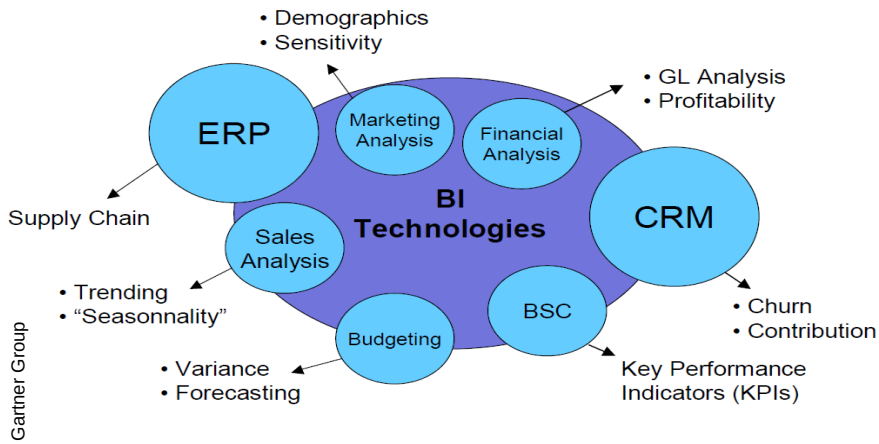
DW - DW vs. Meta-Arquitetura Organizacional



16

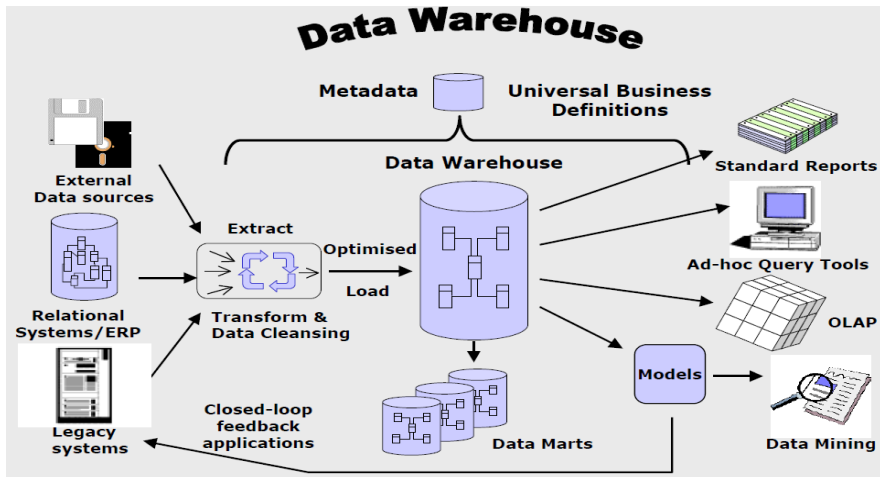
DW – Aplicações vs.. Business Intelligence

- **Application Intersection With Business Intelligence (BI)**



17

DW - Soma de Componentes ou Conceito



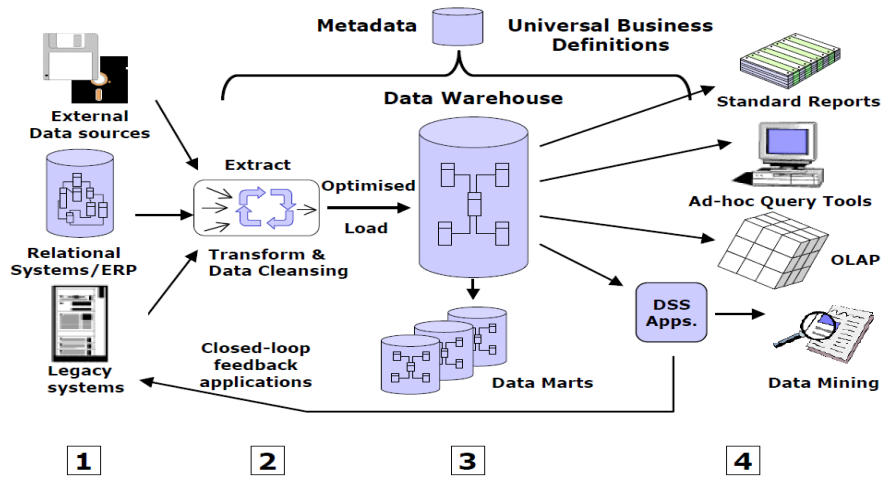
18

DW – Níveis da Arquitetura (1)

- **São 4 os níveis da Arquitetura de um DW:**
  1. Fontes Operacionais de Dados
  2. Área de Retenção (*Data Staging Area*)
  3. Ambiente de Apresentação – "o Data Warehouse"
  4. Aplicações para o Utilizador

19

DW – Níveis da Arquitetura (2)



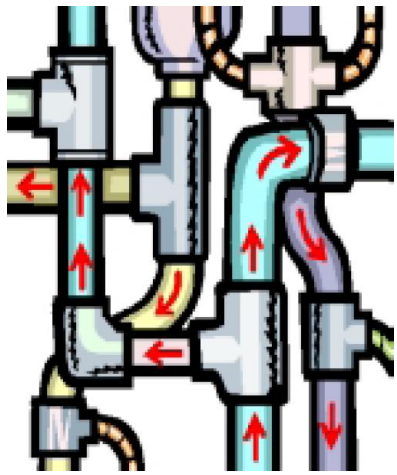
20

DW – Fontes Operacionais (1)

- Críticas do ponto de vista operacional (Exemplos: EPOS num Retalhista ou Billing System numa MEO/Altice)
- Disponibilidade do sistema é a prioridade (Exemplos: na hora de expediente até 24 horas por dia, 7 vezes por semana)
- Dados históricos e capacidades de reporting limitadas ou apenas orientados a objetivos operacionais (Exemplos: Imprimir um ticket, um contrato ou uma fatura)
- Pouco esforço de uniformização de dimensões nas diferentes soluções operacionais (Exemplos: Produto, Cliente, Fornecedor, Calendário)
- Fixação por eleição de atributos para chaves de acesso aos dados – *production keys* (Exemplos: Número de Cliente, Código de Produto)

21

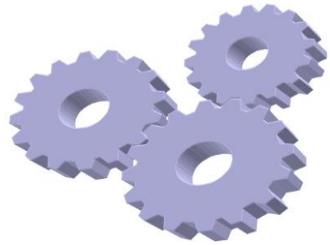
DW – Fontes Operacionais (2)



- Os dados das fontes operacionais são canalizados para o DW
- Parte significativa do trabalho de montar um DW é de canalização
- Quanto maior o volume de dados e a frequência de alimentação do DW mais robusta tem de ser a canalização
- A diversidade das fontes operacionais introduzem complexidade adicional no envio dos dados para o DW

22

DW – Área de Retenção (*Data Staging Area*)



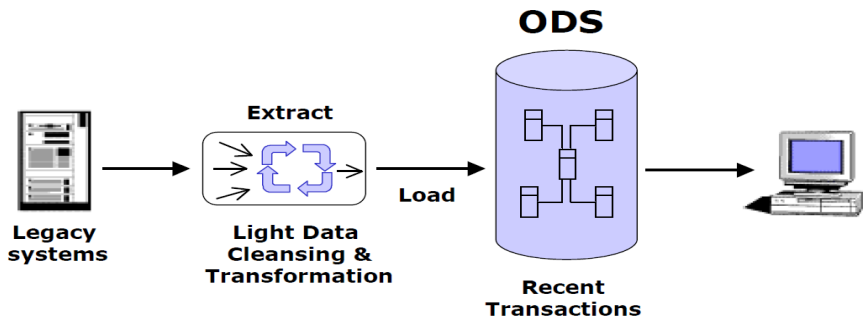
- Área de armazenamento temporário (estágio)
- Conjunto de processos para:
  - Limpar
  - Transformar
  - Combinar
  - Desmultiplicar
  - Validar ou introduzir semântica
  - Arquivar (backup) e preparar os dados para serem usados no DW

- Baseada usualmente em tecnologia relacional (não é pré-condição)
- Não suporta interrogação (*queries*) nem serviços de apresentação (restrição mais importante)

DW – DSA vs. ODS

- **Data Staging ≠ Area Operacional Data Store**
  - Não é um complemento aos sistemas operacionais, nem às dificuldades de manipular informação nos *legacy systems*
  - Não é implementada necessariamente numa base de dados relacional
  - Não replica os modelos de dados operacionais (E&R) complementados com *time stamping*
  - Não serve como área de retenção permanente (Estágio = atividade de duração curta)
  - Serve as funções de carregamento, limpeza, combinação, arquivo (eventualmente), e publicação para exploração

DW – Operacional Data Store (*ODS*)

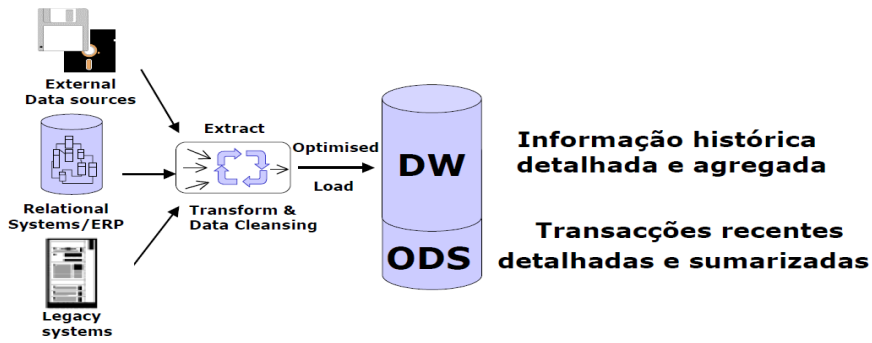


- **Conceito original equivalente à definição de sistema operacional – consolidar informação transacional para utilização transacional**

25

DW – Operacional Data Warehouse

- **A evolução tecnológica e a procura de informação modificou o conceito de ODS para ODW**



26

DW - Processos Básicos

- Transformação:
  - Limpeza
  - Reformatação
  - Combinação
- Carregamento
- Controlo de Qualidade
- Publicação
- Atualização/*Refresh*
- Interrogação
- Auditoria
- Backup
- Recuperação (a partir de
- Backup)

Sobre os Dados

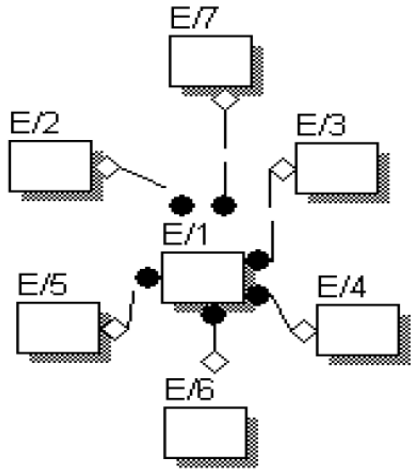
27

DW - Ambiente de Apresentação

- Ambiente onde os dados do DW são organizados e armazenados para interrogação direta pelos utilizadores através de ferramentas de *reporting* ou *query*, ou usada por outros e diversos tipos de aplicações
- É este nível que nos devemos preocupar em aplicar o *framework* dimensional como metodologia de modelação
- Pode ser implementado de 2 formas (ou as duas em combinado):
  - Base de dados relacional - dados organizados em *Star Schemas*
  - Motor OLAP - dados organizados em estruturas dimensionais

28

DW - Modelagem Dimensional



- Modelação alternativa aos modelos de Entidades e Relacionamentos (E&R)
- Modelos compostos por tabelas de:
  - **FACTOS**
- e acompanhadas por tabelas de:
  - **DIMENSÕES**

29

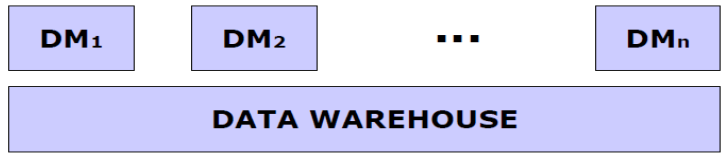
DW – Data Mart

Data Mart = Subconjunto lógico do DW

Do ponto de vista do utilizador

DW = { Data-Marts }

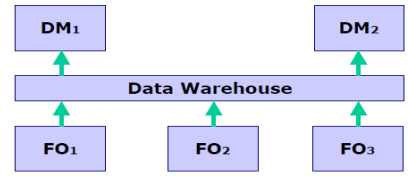
- Cada **Data Mart** é composto por:
  - Um modelo dimensional bem definido
  - Um conjunto de tabelas de factos em que se suporta



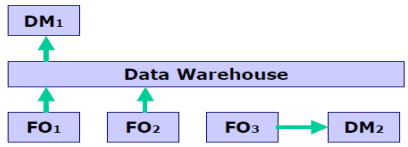
30

DW - Arquiteturas Alternativas de DW

1. DataMart Local



2. "Traditional" Data-Warehouse

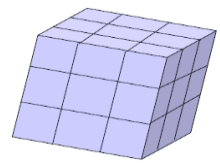


3. Federated Data-Warehouse

31

DW – OLAP: Online Analytic Processing

OLAP = **O**n**L**ine **A**nalytic **P**rocessing



- Forma de organizar a informação e que permite a sua visualização e manipulação sob a forma multidimensional
  - Cubos de dados com n-dimensões
- Baseia-se numa infraestrutura relacional (ROLAP) ou multidimensional (MOLAP) ou hibrida (HOLAP)
  - Representação física dos cubos montados *on-the-fly*

32

DW – OLAP: Online Analytic Processing

MOLAP = **M**ultidimensional **O**n**L**ine **A**nalytic **P**rocessing

- É a forma tradicional de análise.
- A data é armazenada num cubo multidimensional em formato proprietário e não na base de dados relacional.

- **Vantagens**:
  - Uma performance excelente, pois os cubos são construídos para uma extração rápida e ótima em termos de projeção e seleção (*slicing and dicing operations*).
  - Possibilitam a execução de cálculos complexos: todos os cálculos foram pré-gerados com a criação do cubo
- **Desvantagens**:
  - Limitações ao nível do volume de informação a manusear, porque todas os cálculos são executados com a construção do cubo. Não é possível incluir um volume de informação maior no cubo, isto não significa que a informação não possa ser derivada de um volume de data maior, no caso apenas informação sumariada será incluída no cubo.
  - Requer investimentos adicionais: a tecnologia dos cubos é proprietária e não existe na organização. Consequentemente, a adoção da tecnologia MOLAP requer investimentos adicionais em recursos humanos e equipamento.

33

DW – OLAP: Online Analytic Processing

ROLAP = **R**elational **O**n**L**ine **A**nalytic **P**rocessing

- Manipula a data na base de dados relacional
- Cada ação de projeção e seleção (*slicing and dicing operations*) é equivalente ao comando SQL "WHERE"

- **Vantagens**:
  - Lida com grandes volumes de dados: a limitação é o tamanho dos dados na base de dados relacional
  - Aproveita muitas das funcionalidades que vêm incluídas geralmente na DBMS
- **Desvantagens**:
  - O desempenho pode ser lento, pois cada relatório ROLAP é essencialmente uma consulta SQL (ou várias consultas SQL) à base de dados relacional
  - Limitado pelas funcionalidades do SQL e não se ajustam a todas as necessidades (por exemplo, é difícil para realizar cálculos complexos, utilizando SQL)
  - Fornecedores ROLAP têm ultrapassado essas dificuldades com a construção de ferramentas *out-of-the-box para* funções complexas

34

DW - Ad Hoc Query Tool

Utlizadores com Front-Ends Pré-Formatados	Utlizadores sem Front-Ends Pré-Formatados
90% dos Queries	10% dos Queries
OLAP + Queries Pré-Desenhados	Queries desenhados no momento

- *Awareness* para o peso da computação sobre o DW é fundamental!
- Eventual ativação de *features* como pré-requisito do custo do Query

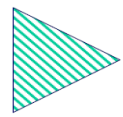
35

DSS - Classes de DSS

DSS = Decision Support System

DSS
- Passivo → Suporta decisões de utilizadores
- Activo → Automa (total ou parcialmente) decisões de utilizadores

Conectividade
Velocidade
Conteúdos



Emergência de um cada vez maior nº de aplicações de DSS Activo

36

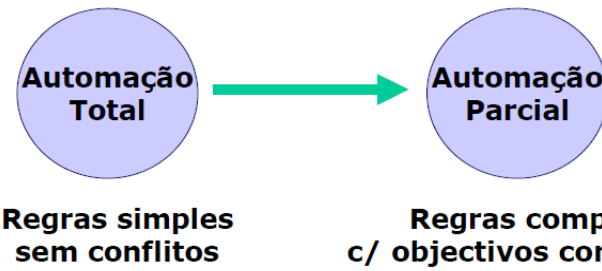
DSS - DSS Passivo

Aplicações de *Queries* AD-HOC	Melhor visibilidade Melhor Suporte às Decisões Melhores Decisões?
Aplicação OLAP	Melhor navegabilidade Melhor Suporte às Decisões Melhores Decisões?
Aplicações de Modelagem (Ex: *Forecasting*)	Maior Rigor Possibilidade de Simulação Melhores Decisões (a menos da questão da adequação do modelo)

Continua a ser perfeitamente crítico o elemento humano de julgamento na decisão !

37

DSS - DSS Ativo



Regras simples sem conflitos

Regras complexas c/ objectivos conflituantes

- Sistemas periciais são a base do domínio de DSS Ativos
- Paradigma fundamental: representação do conhecimento
- Quanto mais alto é o nível de perícia, maior é a percentagem da decisão que deve ser deixada ao homem
- O DW só constitui a base de visibilidade que suporta este tipo de aplicações de *Business Intelligence*

38

DSS - DSS Híbrido

Data Mining

Não parte do paradigma de verificação de hipóteses
Baseia-se exclusivamente nos dados

- É ativo no *Kernel* (técnicas de Data Mining)
- É passivo nas fases de preparação dos dados e interpretação dos dados
- Ainda muito que fazer em termos de usabilidade neste domínio

39

DSS - Exemplos de Aplicações de Data Mining

Projecto Falcon

HNC Software
Pré-Profiling de Clientes (Amex, Visa, Mastercard)
Tempo de intervenção: segundos
Baseado em probabilidades
Redução sensível dos índices de fraude, com menor custo de administração e mais elevada satisfação dos clientes
Capturou 70% do mercado (Bancos de Retalho)

Double Click.com

Empresa de 1 pessoa
"Broker" de "Cookies"
Intermediário entre a oferta de publicidade na NET (lojas virtuais) e os consumidores
Idéia-chave: os "Banners" certos para as pessoas certas
Capturou cerca de 90% do mercado!

40

DSS - Exemplo de Aplicação de DSS Ativo

- **Marcação de PVP no Retalho**

Vendas e Margens	**Shopping Concorrência**
Plano de Margens	**Estratégia de Categorias**
	Posicionamento Competitivo
Resolução de Conflitos	
Decisão de Remarcação PVP	

- Esforço de Modelagem
- Resistência inicial à Mudança
- Piloto de Negócio com Análise de Impacto
- *Roll out* controlado
- Regras diferentes para Negócios diferentes

41

DW – Metadata

***Metadata* (definição estreita) =**
Dados sobre os Dados

Exemplo: o que são Vendas?
(líquidas? ou ilíquidas?)

***Metadata* (definição alargada) =**
Dados sobre os Dados e as transformações que eles passam

Exemplo: documentação completa do processo de extração, limpeza, retenção e agregação

- O conceito de Metadata é muito importante para os projetos de DW, de EAI, e de todo o *middleware* em geral

42

DW – Projeto Data Warehouse (1)

Precisamos de um Data Warehouse ✓

Estamos prontos para o ter ?

- **Temos uma cultura de informação ?**
  - Gerimos por instinto;
  - Não temos falta de dados;
  - Ninguém lê os relatórios existentes;
  - A nossa secretária é que usa o PC por nós;
  - Vivemos obcecados com o secretismo.
- **Temos uma cultura de informação ?**
  - Gerimos com factos;
  - Estamos sempre à procura de novos dados;
  - Redigitamos todos os relatórios em folhas de cálculo;
  - O nosso PC é-nos indispensável;
  - Partilhamos a informação.

43

DW – Projeto Data Warehouse (2)

Precisamos de um Data Warehouse ✓

Estamos prontos para o ter ✓

Podemos construí-lo ?

- **Temos uma cultura de Tecnologia?**
  - Temos de comprar todo o hardware, middleware e software;
  - Estamos ocupados a consolidar empresas, a resolver, a implementar SAP;
  - Estamos à espera que o novo ERP resolva o problema da consistência de dados.
- **Temos uma cultura de Tecnologia?**
  - Temos uma infraestrutura;
  - Temos uma equipa disponível;
  - Temos dados fiáveis e acreditamos que a publicação de dados menos fiáveis pode motivar uma reengenharia de processos.

44

DW – Projeto Data Warehouse (3)

Precisamos de um Data Warehouse ✓

Estamos prontos para o ter ✓

Podemos construí-lo ✓

Podemos pagá-lo ?

- **Há muitas histórias de ROI ...**
- Procure um bom Sponsor !

45

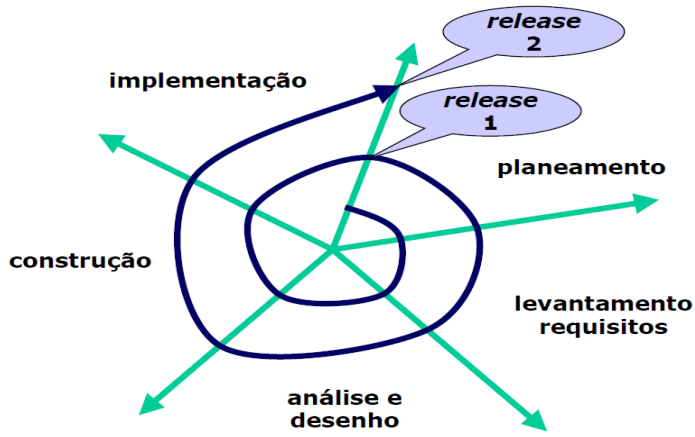
DW – Fatores Críticos de Sucesso

- **Sponsorização de Topo**
- **Motivação do Negócio**
- **Viabilidade (dados, tecnologia)**
- **Técnicas de Modelagem**
- **Parcerias Tecnológicas**
- **Cultura Analítica**

46

DW – Ciclo de Desenvolvimento (1)

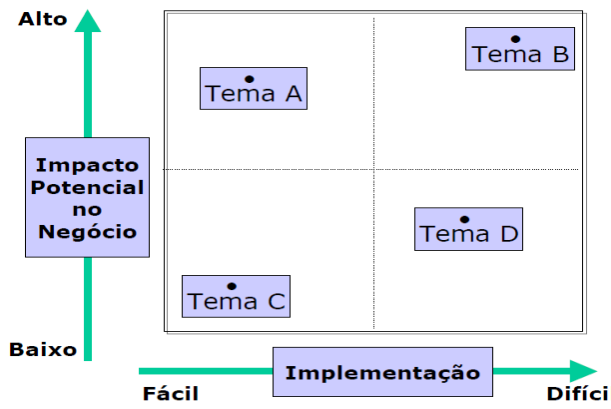
- **Não vamos fazer tudo de uma só vez !**



47

DW – Ciclo de Desenvolvimento (2)

- **Escolher bem o que fazer !**

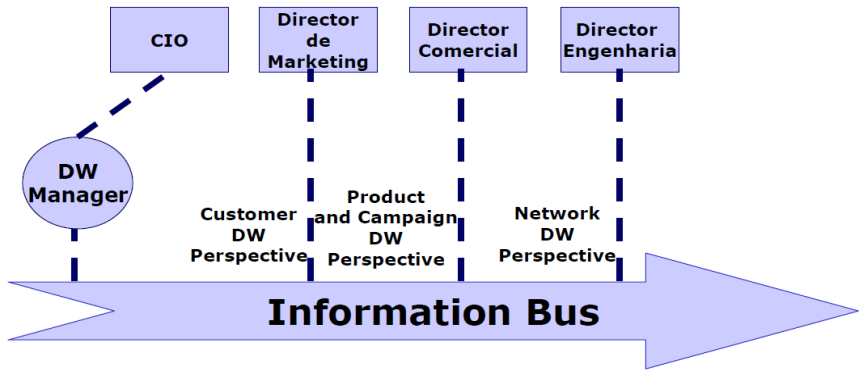


Análise por quadrantes - Kimball

48

DW – Gerir um Data Warehouse

- **O DW Manager na Organização**



49

DW - DW Manager: Definição da Função

- "Guardião do Templo"
- "Editor" responsável pela qualidade dos dados publicados
- Responsável pela Metadata Organizacional
- Facilitador da priorização do desenvolvimento de todas as aplicações de Suporte à Decisão
- Responsável pela publicação, nos timings acordados, das novas versões dos dados

50

DW – DW Manager: Valências Exigidas

- Conhecimento do Negócio (em particular, das necessidades dos *Knowledge Workers*)
- Capacidade de comunicação e facilitação / geração de consenso
- Capacidade de organização disciplina de entrega (c/ controlo de qualidade)
- Conhecimentos técnicos específicos de DW / DSS / BI
- Resiliência (DW é um Processo!)
- Capacidade de modelagem avançada

51

IT - Dimensão de um DW

Qual o tamanho da "Piscina" ?

Dados
- **Números de Tabelas**
- **Volume de Dados**
- **Número de Fontes operacionais (instâncias)**
- **Número de *Data Marts***

Processos
- **Número de Processos *batch***
- **Número de *Queries* por dia**

Quantos vão mergulhar ?

IMPACTO NA INFRAESTRUTURA !

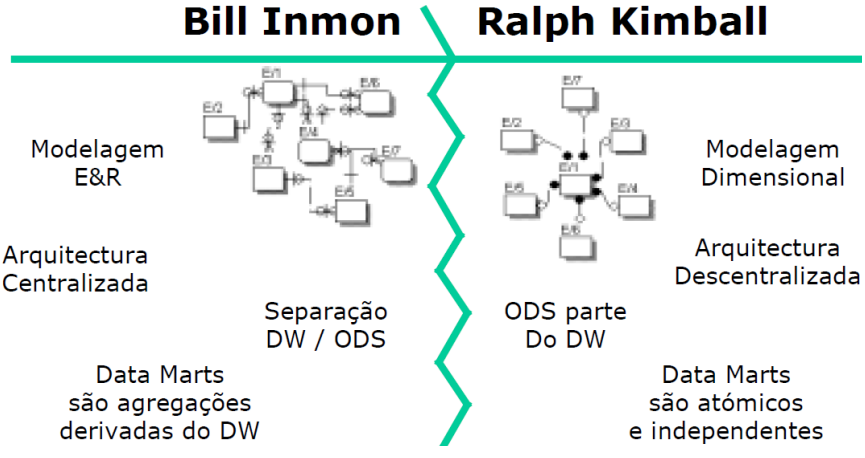
52

IT - Problemas técnicos que se levantam

- Como facilitar a paralelização implícita do software?
- Distribuição da computação SMP ou MPP?
- Como promover o balanceamento dinâmico de carga entre processadores?
- Como diminuir ao máximo a contenção entre processos?
- Como garantir que as operações pouco extensas de *delete*/*update* não têm grande impacto de performance?
- Como garantir *backups* íntegros e prontos a suportar uma recuperação?
- Como implementar *Disaster Recovery*?

53

DW – DW Gurus e suas Metodologias (1)



54

DW – DW Gurus e suas Metodologias (2)

- Bill Inmon vs. Ralph Kimball
  - http://www.1keydata.com/datawarehousing/inmon-kimball.html
- Kimball vs. Inmon...or, How to build a Data Warehouse
  - http://it.toolbox.com/blogs/confessions/kimball-vs-inmonor-how-to-build-a-data-warehouse-10987
- Tutorial 4 : Design of the data warehouse: Kimball Vs Inmon
  - http://www.exforsys.com/tutorials/msas/data-warehouse-design-kimball-vs-inmon.html
- Inmon vs.. Kimball - An Analysis
  - http://www.nagesh.com/publications/technology/173-inmon-vs-kimball-an-analysis.html
- **Abordagem Kimball Vs. Abordagem Inmon**
  - http://www.slideshare.net/guest2308b5/kimball-vs-inmon
- **Microsoft Business Intelligence de Ponta-a-Ponta**
  - http://msdn.microsoft.com/pt-br/library/cc517991.aspx

Todos os links acedidos 18/set/2018