

USP

LABIC

Navegando nos dados

Data Warehouse e OLAP

Solange Oliveira Rezende

USP São Carlos / ICMC
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
<http://www.icmc.usp.br>

© Solange Oliveira Rezende

USP

LABIC

Evolução das Tecnologia Relacionadas com Dados

EVOLUÇÃO	TECNOLOGIA UTILIZADA
Coleta de dados (1960s)	Computadores, Fitas, Discos
Acesso aos Dados (1980s)	RDBMS, SQL, ODBC
Navegação pelos dados (1990s)	SGBD, OLAP, Base de Dados Multidimensionais, Data Warehouse
Data Mining (2000)	Algoritmos Avançados, Computadores com Multiprocessadores, Grandes Bases de Dados

SOF

USP

LABIC

Motivação

Passado

- Tecnologia limitada
- Armazenamento de pequenos volumes de dados (Mbytes)
- Consultas aos dados
- Não existiam ferramentas para auxiliar a análise das informações obtidas

Presente/Futuro

- Grandes avanços tecnológicos na área de Tecnologia de Informação
- Armazenamento de grandes volumes de dados (Tbytes)
- Necessidade de conhecer e entender a BD
- O conhecimento extraído de uma BD deve ser usado para auxiliar as tomadas de decisões.

SOF

USP

LABIC

Motivação (cont.)

Pirâmide do Conhecimento

The diagram illustrates the Pyramid of Knowledge with three levels: DADO (Data) at the base, INFORMAÇÃO (Information) in the middle, and CONHECIMENTO (Knowledge) at the top. To the left, a vertical arrow indicates the progression from 'Consultas à BD' (Database Queries) at the bottom to 'Utilização do Conhecimento' (Knowledge Utilization) at the top. To the right, another vertical arrow shows the progression from 'Passado' (Past) at the bottom to 'Presente/Futuro' (Present/Future) at the top. The middle level is labeled 'Obtenção do Conhecimento' (Knowledge Acquisition).

SOF

USP

LABIC

Introdução

Atualmente, empresas e organizações estão interessadas em utilizar todas as informações que dispõem para tomar as melhores decisões.

SOF

USP

LABIC

Introdução

A maioria dos bancos de dados operacionais (BD) utilizados nessas empresas não estão adequados.

SOF

USP **Introdução** LABC USP NUMA LABIC

Dificuldades encontradas em um BD quando utilizados para auxiliar na tomada de decisão

Dados:

- ✧ não apresentam um contexto histórico
- ✧ podem não estar padronizados para serem analisados conjuntamente


Consultas:

- ✧ podem exigir grandes esforços
- ✧ demandam muito tempo de processamento

SOF

USP **Introdução** LABC USP NUMA LABIC

Solução para as dificuldades:



Integrar os dados da empresa/organização em uma estrutura única que permita uma melhor e mais eficiente utilização dos dados.

SOF

USP **Data Warehousing** LABIC

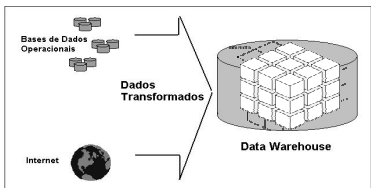
Definição

- ✧ *Data Warehousing* é um processo, não um produto, para montar e gerenciar dados de várias fontes com o propósito de ganhar uma visão detalhada e singular de parte ou do todo de um negócio
- ✧ O produto gerado de um projeto de *Data Warehousing* é o seu *Data Warehouse* (DW)

SOF

USP **Data Warehouse** LABIC

- O objetivo principal é ter uma visão mais ampla das informações relacionadas à empresa/organização.
- É responsável pelo agrupamento dos dados históricos da empresa



SOF

USP **O que é um Data Warehouse?** LABIC

É uma coleção de dados com as seguintes características:

- É baseado em assuntos
- É integrado
- É não-volátil
- Apresenta um contexto histórico

SOF

USP **O que é um Data Warehouse?** LABIC

É uma coleção de dados com as seguintes características:

- É baseado em assuntos

Os dados devem ser orientados ou organizados por assunto para que as informações possam ser mais direcionadas à tomada de decisão.

W. H. Inmon

SOF

USP

LABIO

O que é um Data Warehouse?

É uma coleção de dados com as seguintes características:

- É baseado em assuntos
- É integrado

As inconsistências dos dados extraídos devem ser desfeitas.

W. H. Inmon

SOF

USP

LABIO

O que é um Data Warehouse?

É uma coleção de dados com as seguintes características:

- É baseado em assuntos
- É integrado
- É não-volátil

Não sofrem modificações, sendo somente carregados e acessados.

W. H. Inmon

SOF

USP

LABIO

O que é um Data Warehouse?

É uma coleção de dados com as seguintes características:

Os dados históricos são mantidos, ou seja, dados sobre um determinado assunto, tratados em diferentes tempos, são armazenados.

- Apresenta um contexto histórico

W. H. Inmon

SOF

USP

LABIO

O que é um Data Warehouse?

Outros dois aspectos de um DW devem ser considerados:

- **Particionamento dos Dados:** dividir os dados em mais de uma unidade física para proporcionar maior flexibilidade no gerenciamento.
- **Granularidade:** se refere ao nível de detalhamento dos dados.

SOF

USP

LABIO

Granularidade

Data Warehouse

Nível de detalhe resumido

Nível de detalhe sem resumo

Banco de Dados

Banco de Dados Operacionais

Granularidade

+

↑

-

SOF

USP

LABIO

BD versus DW

Bases de Dados Operacionais e Data Warehouse apresentam características diferentes.

SOF

USP

LABIO

BD versus DW

- **Base de Dados Operacionais:** utilizam sistemas de processamento de transação on-line (OLTP).
- **Data Warehouse:** utilizam técnicas de processamento analítico on-line (OLAP).

SOF

USP

LABIO

BD versus DW

Características	BD	DW
Dados	Atuais, Isolados, Relacionais	Históricos, Integrados, Resumidos
Probabilidade de Acesso	Alta	Baixa, Moderada
Usuários	Adm. Sistemas Projetista Sistema Operadores	Analista, Executivo e Usuários do Conhecimento
Modelagem	MER	Dimensional
Projeto	Orientado à Aplicação	Orientado a Assunto, Negócio
Processamento	Repetitivo	"Ad-hoc"

SOF

USP

LABIO

Data Warehouse

- ♦ **Topologias:**
 - ✧ Centralizada
 - ✧ Data Mart
 - Dependentes
 - Independentes
 - ✧ Distribuída
- ♦ **Metadados:**
 - ✧ Dados sobre os dados
 - ✧ Abstração dos Dados

SOF

USP

LABIO

USP

LABIO

Projeto de um Data Warehouse

- ♦ Entendimento do Domínio da Aplicação
- ♦ Coleta, Transformação e Integração de Dados
- ♦ Modelagem do Data Warehouse
- ♦ Gerenciamento do Data Warehouse

SOF

USP

LABIO

USP

LABIO

Projeto de um Data Warehouse

Para se construir um Data Warehouse, dois aspectos importantes devem ser considerados:

- ✧ A forma como irá modelar e armazenar os dados
- ✧ Quais métodos de extração e integração de dados serão utilizados

SOF

USP

LABIO

USP

LABIO

Extração, Transformação e Integração

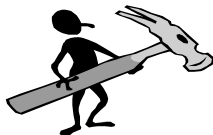
Para que uma extração ocorra, é necessário que um programa:

- ✧ Percorra um arquivo ou BD fonte
- ✧ Utilize critérios de seleção
- ✧ Transporte os dados selecionados para outro arquivo ou BD.

SOF


Extração, Transformação e Integração

Os dados extraídos devem, muitas vezes, passar por transformações para que as inconsistências sejam desfeitas e que todos os dados apresentem o mesmo formato no DW.



Extração, Transformação e Integração

Os dados devem ser integrados e terem um elemento de tempo vinculados a eles para que uma visão mais abrangente das informações possam ser obtidas.



Modelagem de Dados

A modelagem dos dados é utilizada para a estruturação dos dados para permitir que consultas, considerando diferentes aspectos de negócios, possam ser realizadas.

Uma técnica de modelagem de dados conhecida é a Modelagem Dimensional.

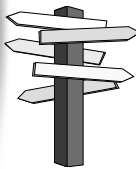
Modelagem de Dados

Na Modelagem Dimensional

- Os dados são mapeados em uma estrutura com mais de uma dimensão para modelar os dados
- As dimensões da estrutura podem representar, por exemplo, os negócios de uma empresa
- As medidas são os dados encontrados através de consultas

Modelagem de Dados

Existem vários esquemas que mantêm a multidimensionalidade dos dados e que podem representar os dados em um DW:



- Modelagem Multidimensional
- Esquema Star
- Esquema Snowflake

Modelo Multidimensional

- Os dados são manipulados como hipercubos
- Os dados são mapeados em arranjos n-dimensionais
- O acesso aos dados geralmente é rápido e fácil

Modelo Multidimensional		
Exemplo de um Modelo Relacional		
Produto	Localização	Vendas
Automóvel	MG	200
Automóvel	SP	300
Automóvel	RS	150
Motocicleta	MG	63
Motocicleta	SP	78
Motocicleta	RS	72
Pick-up	MG	30
Pick-up	SP	56
Pick-up	RS	24

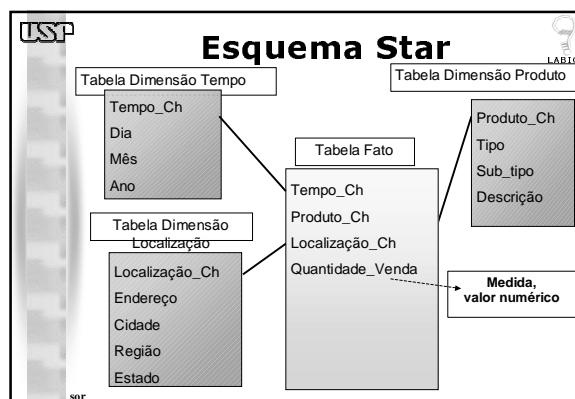
Modelo Multidimensional			
Na tecnologia multidimensional			
	MG	SP	RS
Automóvel	200	300	150
Motocicleta	30	56	24
Pick-up	63	78	72

Vantagens: menor espaço de armazenamento
mais intuitivo para modelagem dimensional

- ### Esquema Star
- No Esquema Star existem dois tipos de tabela:
- Tabela Fato
 - Tabela de Dimensão

- ### Esquema Star
- No Esquema Star existem dois tipos de tabela:
- Tabela Fato
- É constituída de campos que representam:
- as medidas
 - as tabelas de dimensão

- ### Esquema Star
- No Esquema Star existem dois tipos de tabela:
- Representa cada dimensão do modelo
- contém informação textual para descrever as medidas
 - não é normalizada



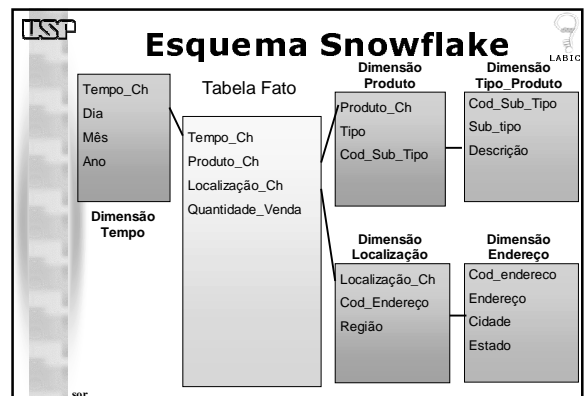
USP

LABIC

Esquema Snowflake

- Possui tabelas fato e de dimensão
- As tabelas de dimensão são normalizadas:
 - ↗ as redundâncias e os campos nulos são eliminados
 - ↗ o espaço físico é economizado

SOF



USP

LABIC

Componentes de um DW

São componentes de um DW:

- Metadados
- Cubo
- Ferramentas de extração

SOF

USP

LABIC

Componentes de um DW

São componentes de um DW:

- Metadados

São dados sobre os dados, ou seja, é a chave para entender o conteúdo e a estrutura de um DW
- Ferramentas de extração

SOF

USP

LABIC

Componentes de um DW

São componentes de um DW:

- Metadados
- Cubo

É um conjunto de células de dados combinados pelas diferentes dimensões

SOF

USP

LABIC

Componentes de um DW

São componentes de um DW:

- Metadados

Extraem, selecionam, integram e carregam os dados das fontes para o Data Warehouse

SOF


USP

LABIO

Metadado

Permite que o analista tenha informações sobre o que procurar e onde encontrar no DW.

Assim, o analista pode encaminhar suas consultas e desempenhar sua tarefa eficientemente.



SOF

USP

LABIO

Metadados

Normalmente, os metadados têm informações sobre os seguintes aspectos:

- A estrutura dos dados segundo a visão do analista
- A estrutura dos dados segundo a visão do programador
- O modelo de dados

SOF

USP

LABIO

Metadados (cont.)

- A fonte da qual os dados são retirados
- O histórico das extrações
- A transformação pela qual passaram os dados
- As estruturas das tabelas do DW
- Os atributos das tabelas do DW
- As rotinas comuns de acesso

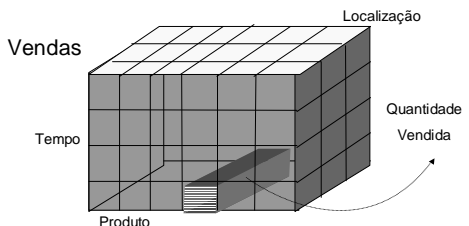
SOF

USP

LABIO

Cubo

É a forma de organizar os dados para realizar as consultas, utilizando a modelagem dimensional.

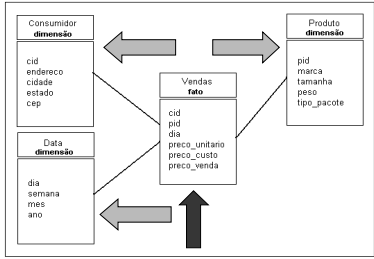


SOF

USP

LABIO

Modelagem Multidimensional

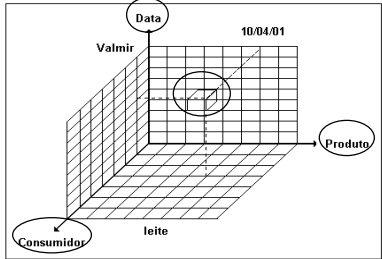


SOF

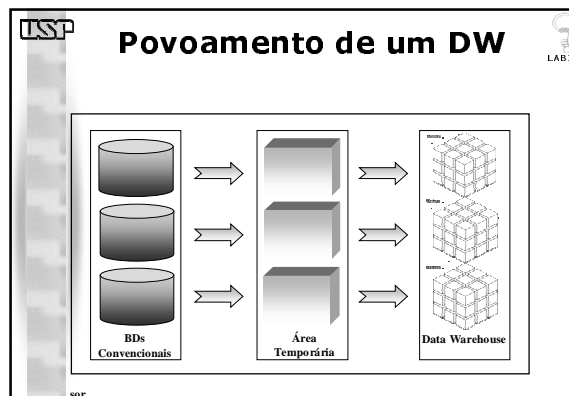
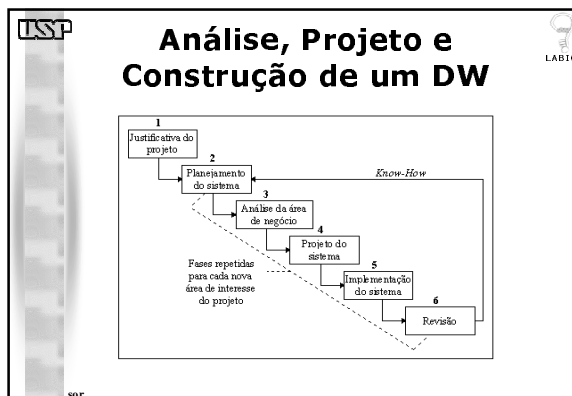
USP

LABIO

Modelagem Multidimensional



SOF



USP **LABIO**

Problemas que Data Warehouse soluciona

- Os Sistemas de Informação Executiva (SIE) foram soluções populares nos anos 80's.
- Os SIE eram simples e pouco flexíveis, mas não permitiam uma busca rápida e com detalhe.

SOF

USP **LABIO**

Problemas que Data Warehouse apoia

- Sistemas de Suporte à Tomada de Decisões e Base de Dados Multidimensionais tinham uma capacidade limitada e um alto grau de manutenção.
- Crescimento desorganizado das grandes bases de dados

SOF

USP **LABIO**

Data Warehouse está dirigido para:

- Empresas que possuem mais de 100 GBytes de dados armazenados.
- Empresas que precisam de uma organização na estrutura de sua informação.
- Empresas enormes que precisam de informações imediatas para a tomada de decisão

SOF

USP **LABIO**

OLAP (On-Line Analytical Processing)

- Voltadas para análise multidimensional de dados de modo superior aos mecanismos oferecidos pelas ferramentas tradicionais


É a análise, síntese e consolidação de grandes volumes de dados multidimensionais [Codd 93].

- Ferramenta geralmente utilizada para a análise de Data Warehouse

SOF

USP **SQL e OLAP (On-Line Analytical Processing)** LABIO

- SQL normal
 - ✧ O usuário/analista tem que ter suporte e compreender SGBD.
- Ferramenta OLAP
 - ✧ É voltada para gerentes. Os resultados geralmente são mostrados em relatórios mais sofisticados mas sem a necessidade de conhecimentos prévios de SGBD.
 - ✧ Apresenta dados relacionais de forma a facilitar a compreensão dos dados.



SOF

USP **Consultas OLAP** LABIO


- Auxiliam os usuários a sintetizar as informações através de visões comparativas e personalizadas, assim como analisar dados históricos.
- É uma tecnologia que possibilita aos usuários acesso:
 - ✧ rápido
 - ✧ consistente
 - ✧ interativo

SOF

USP **Consultas OLAP** LABIO

Pode-se ter diversos tipos de consulta, dentre elas:

- Pivot
- Roll-up
- Slice
- Drill-down/up
- Drill across



SOF

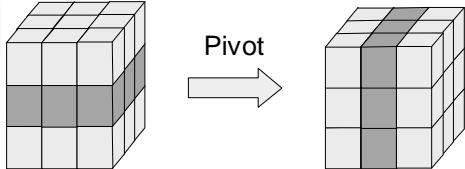
USP **Consultas OLAP** LABIO

- Roll-up: é computar todas as relações de dados para uma ou mais dimensões.
- Drill-across: é o processo de ligar duas ou mais tabelas fato de mesmo nível de detalhes.

SOF

USP **Pivot** LABIO

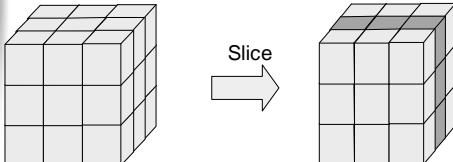
É usado para mudar a orientação dimensional de uma pesquisa.



SOF

USP **Slice** LABIO

Um slice é um subconjunto da estrutura multidimensional que corresponde a um valor simples em lugar de um ou mais atributos das dimensões.



SOF