Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

- Definido de várias maneiras diferentes, mas não de uma forma rigorosa.
  - Uma base dados de suporte a decisão que é mantida separadamente da base operacional da organização.
  - Suporta processamento de informação fornecendo uma plataforma sólida para análise de dados históricos, consolidados.
- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—W. H. Inmon
- Data warehousing:
  O processo de construir e usar data warehouses

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Características que qualificam um DW. (Kimbal)

- a orientação,
- a grande quantidade de dados,
- os fins da organização,
- a integração,
- a informação temporal,
- a não volatilidade,
- as estruturas de dados optimizadas e
- a granularidade.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Orientado por Tema

- Organizado em torno de temas importantes, tais como cliente, produto, vendas.
- Focado na modelação e análise de dados para quem toma decisões, em vez de operações diárias e processamento de transacções.
- Fornece uma visão **simples e concisa** sobre questões de um tema particular através da **exclusão de dados que não são importantes** no suporte ao processo de decisão.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Integrado

- Construído por integração de múltiplas e heterogéneas fontes de dados
  - Bases de dados relacionais, ficheiros simples, registos de transacções on-line
  - São aplicadas técnicas de limpeza de dados e integração de dados.
  - É assegurada a consistência na convenção de nomes, codificação de estruturas, atributos de medidas, etc. entre diferentes fontes de dados
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - Quando a informação é movida para o warehouse, é feita a conversão.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Variável Tempo

- O horizonte de tempo para um data warehouse é significativamente maior do que o de sistemas operacionais.
  - Base de dados operacional: informação actual.
  - Dados no data warehouse: fornece informação numa perspectiva histórica (e.g., últimos 5-10 anos)

- Cada estrutura chave no data warehouse
  - Contém um elemento de tempo, explicita ou implicitamente
  - Mas a chave de dados operacionais pode ou não conter um “elemento de tempo”.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Data Warehouse vs. SGBD Heterogéneos

- Integração tradicional de BD heterogéneas:
  - Construir **conversores/mediadores** sobre BD heterogéneas
  - Abordagem **orientada-a-consulta**
    - Quando uma consulta é feita a uma determinada BD, usa-se um meta-dicionário para traduzir a consulta em consultas apropriadas para outras BD's envolvidas, e os resultados são integrados num conjunto resposta global
    - Filtragem de informação complexa, competição por recursos
- Data warehouse: **orientada-por-actualizaçao**, alta performance
  - A informação de fontes heterogéneas é previamente integrada e guardada em warehouses para consulta e análise directa.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Data Warehouse vs. SGBD Heterogéneos

- OLTP (on-line transaction processing)
  - Tarefa principal dos SGBD relacionais tradicionais
  - Operações diárias: diários clínicos, resultados de análises, saldos, produçao, salários, registo, contabilidade, etc.
- OLAP (on-line analytical processing)
  - Tarefa principal de sistemas de data warehouse
  - Análise de dados e tomada de decisões
- Características distintas (OLTP vs. OLAP):
  - Orientação do sistema e utilizador: paciente vs. comunidade
  - Conteúdo dos dados: actuais, detalhados vs. históricos, consolidados
  - Desenho da BD: ER + aplicação vs. estrela + tema
  - Visão: actual, local vs. evolucionária, integrada
  - Padrões de acesso: actualização vs. consultas read-only, complexas

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

BD Operacionais	BD DW
Objectivos operacionais.	Registo histórico.
Acessos de leitura/escrita.	Acessos só de leitura.
Acesso por transacções predefinidas.	Acesso por consultas e relatórios.
Acessos a poucos registos de cada vez.	Muitos registos em cada acesso.
Dados actualizados em tempo-real.	Carregamentos periódicos de dados.
Estrutura optimizada para actualizações.	Estrutura optimizada para consultas.
Event-driven – os processos geram dados.	*Data-driven* – os dados geram respostas.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Características	OLTP	OLAP
Tamanho	GBytes	Giga a TBytes
Origem dos Dados	Interno	Interno e Externo
Actualização	On-Line	Batch
Períodos	Actual	Histórico
Consultas	Previstas	*Ad Hoc*
Actividade	Operacional	Analítica

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Porquê Separar um Data Warehouse?

- Alta performance para ambos os sistemas
  - **SGBD**— optimizados para OLTP: métodos de acesso, indexação, controlo de concorrência, recuperação;
  - **Warehouse**— optimizado para OLAP: consultas OLAP complexas, visões multi-dimensionais, consolidação.
- Funções diferentes e dados diferentes:
  - **Falta de dados:** suporte à decisão requer dados históricos que BD's operacionais tipicamente não mantém
  - **Consolidação de dados**: SD requer consolidação (agregação, sumarização) de dados de fontes heterogéneas
  - **Qualidade de dados**: Fontes diferentes usam tipicamente representações inconsistentes de dados, códigos e formatos que têm de ser reconciliados

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

De Tabelas e Folhas de Cálculo para Cubos de Dados

- Um data warehouse é baseado num **modelo de dados multidimensional** (Kimbal) que vê os dados na forma de um cubo de dados.
- Um cubo de dados, tal como sales, permite que a informação seja modelada e vista em múltiplas dimensões:
  - Tabelas de dimensão, tais como item (item_name, brand, type), ou time(day, week, month, quarter, year), e
  - Tabelas de factos contém medidas (tais como dollars_sold) e chaves externas para cada tabela de dimensão relacionada.
- Na literatura de data warehousing, um cubo n-D é chamado cubóide. O **cubóide 0-D** de topo, que contém o nível mais alto de sumariazação, é chamado **cubóide apex**. O reticulado de **cubóides** forma o cubo de dados.

Análise de Dados

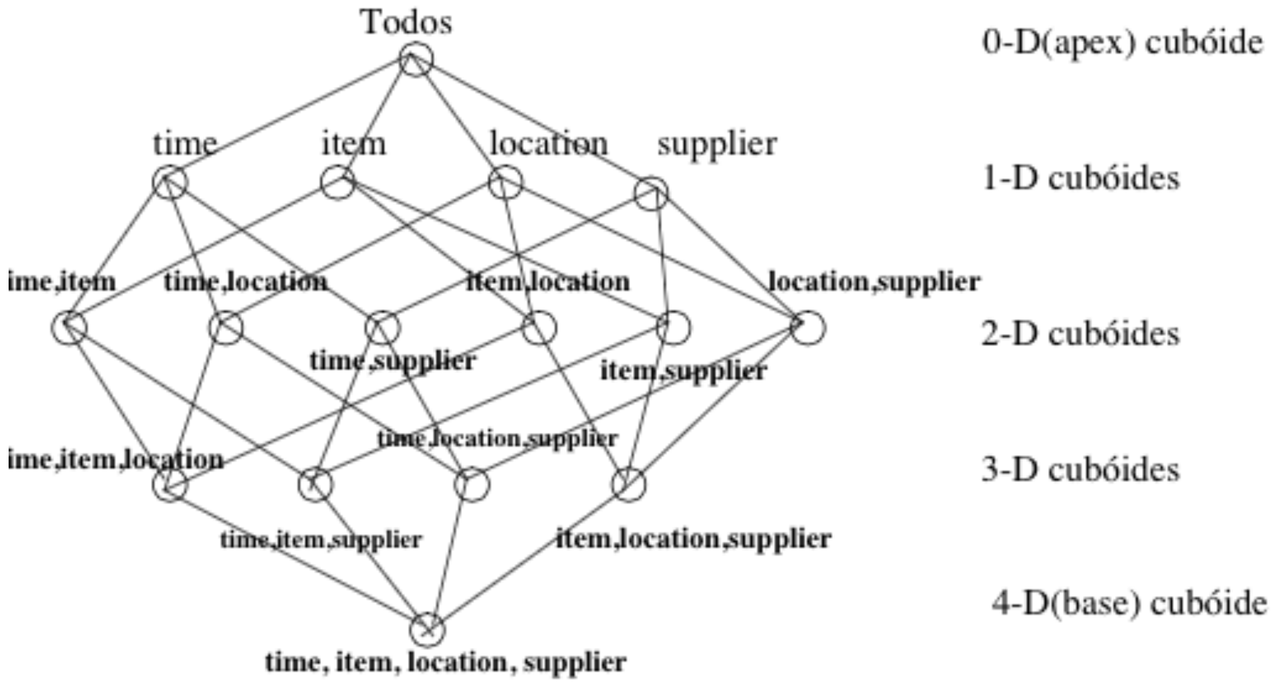
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Cubo: Reticulado de cubóides



Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Modelar data warehouses: dimensões & medidas

- **Esquema estrela**: Tabela de factos no centro ligada a um conjunto de tabelas dimensão
- **Esquema floco de neve**: Um refinamento do esquema estrela onde parte da hierarquia dimensional é normalizada num conjunto de tabelas dimensão mais pequenas, numa forma similar a um floco de neve.
- **Constelações de factos**: Tabelas de factos múltiplas partilham tabelas dimensão, formando um grupo de estrelas, logo chamado constelação de factos.

Análise de Dados

Factos, Dimensões e variáveis

- Conceitos básicos associados a este tipo de modelação:
  - Um **facto** representa um item, uma transacção ou um evento de negócio e é utilizado para analisar o processo de negócio de uma organização.
    - É representado por valores numéricos e
    - implementado em tabelas de factos.
  - As **dimensões** são os elementos que participam num facto, ou seja, as possíveis formas de visualizar os dados.
    - Normalmente não possuem atributos numéricos,
    - descrevem e classificam os elementos que participam num facto.
  - As **variáveis** são os atributos numéricos que representam um facto.
    - São determinadas pela combinação das dimensões que participam de num facto e
    - estão localizadas como atributos de um facto,

Análise de Dados

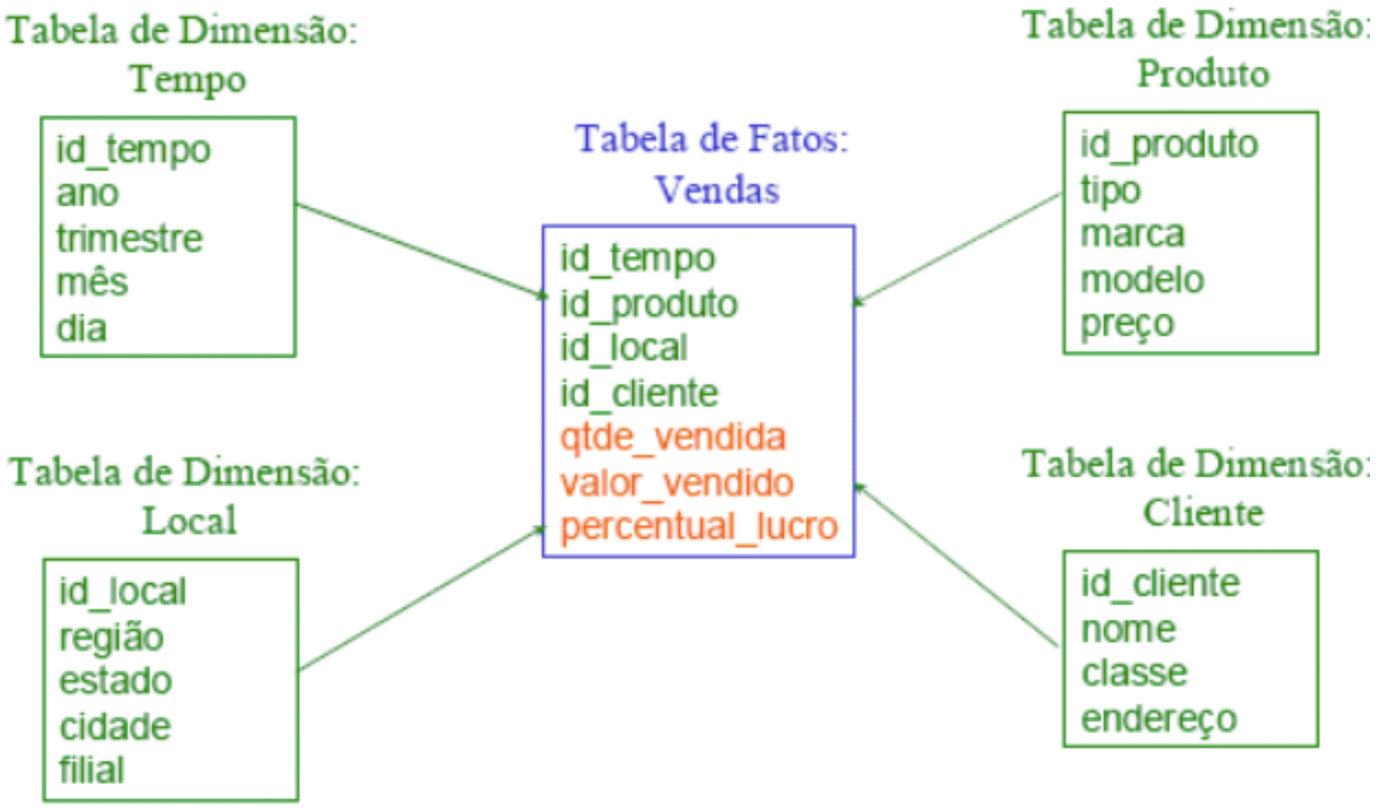
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

- **Esquema estrela (adaptado de Kimbal):**



Análise de Dados

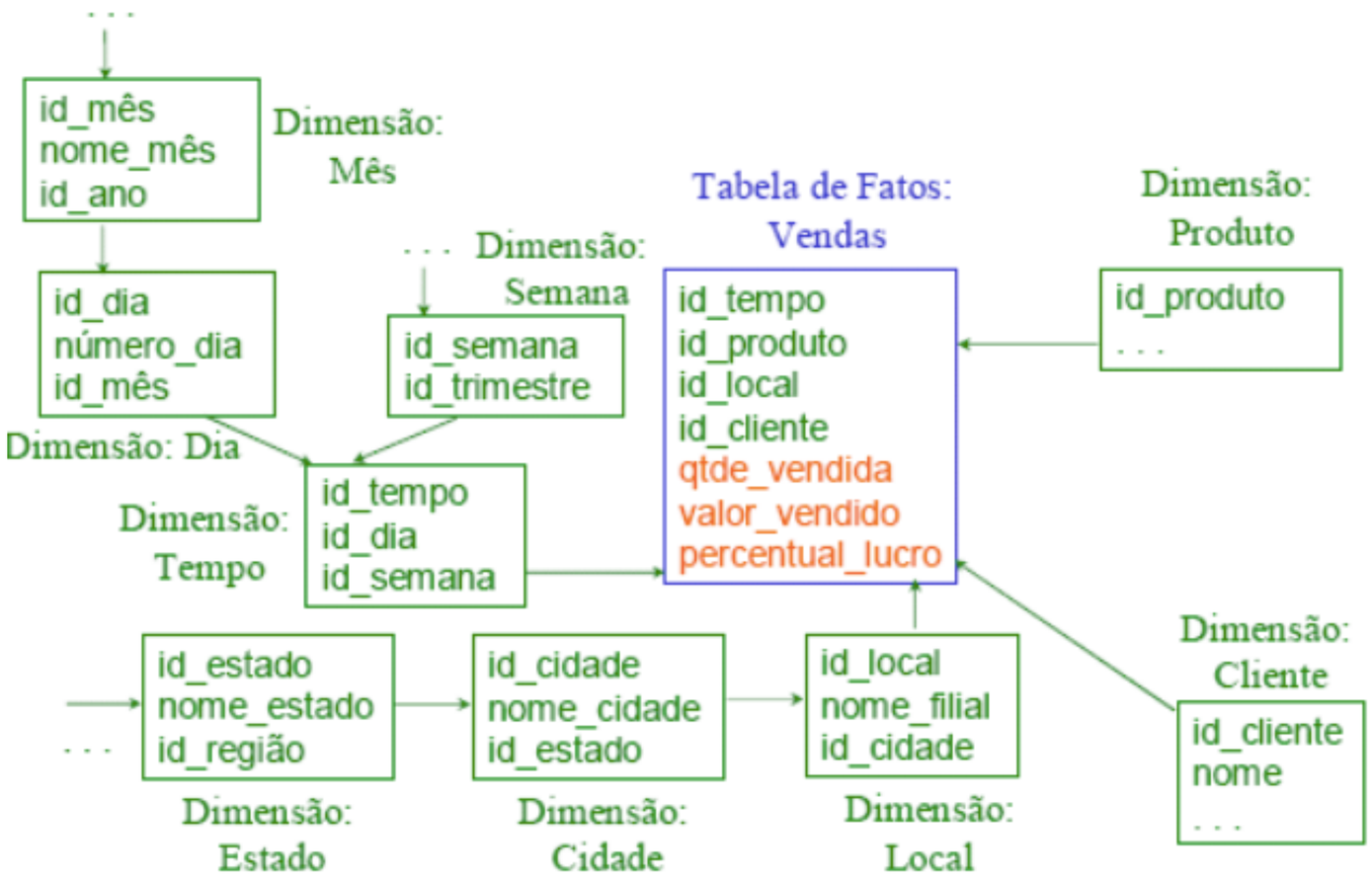
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

- **Esquema floco de neve (adaptado de Kimbal)**:



Análise de Dados

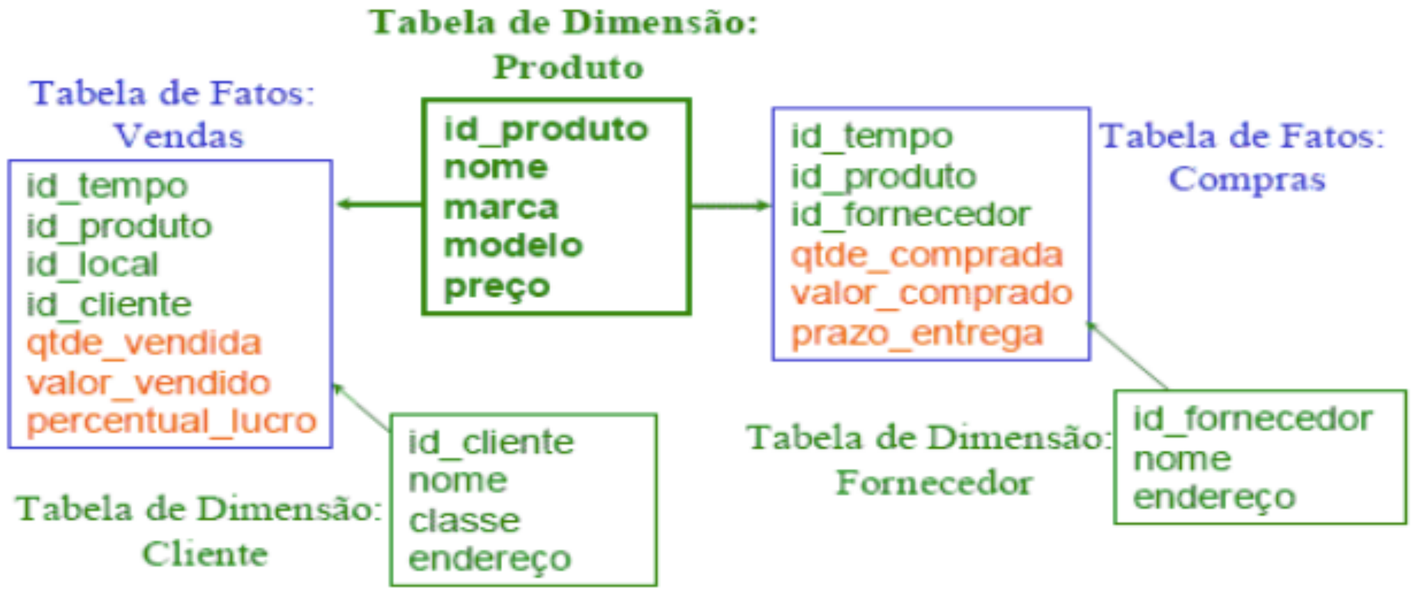
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

- **Constelações de factos (adaptado de Kimbal)**:



Análise de Dados

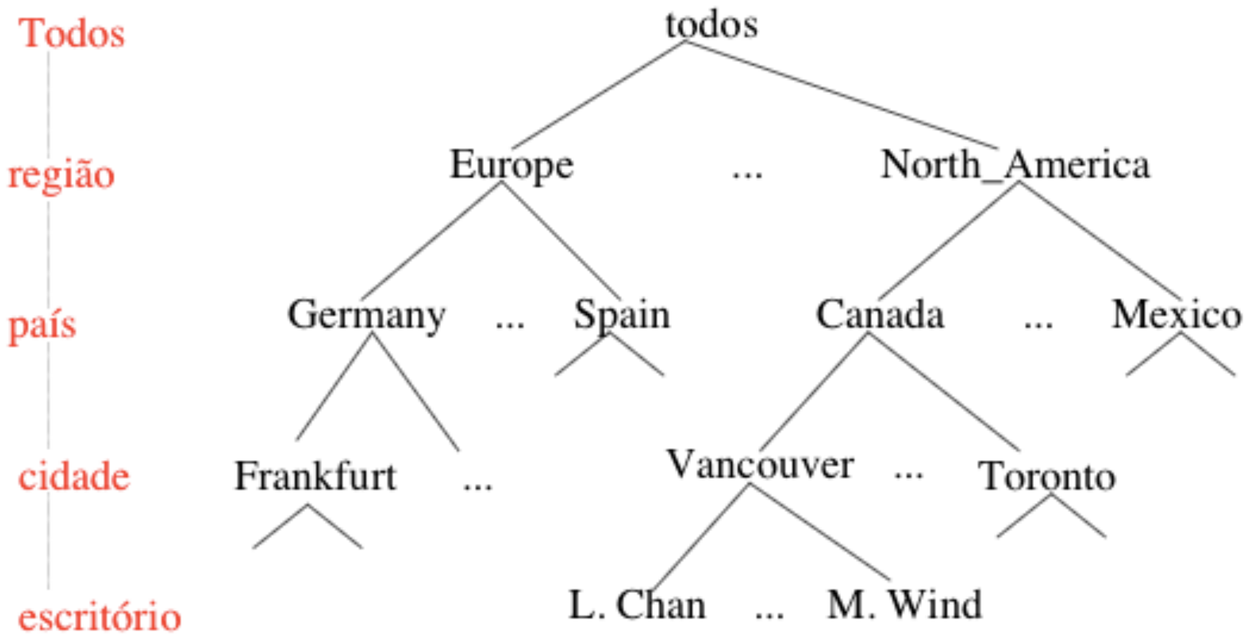
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Hierarquias conceptuais: Dimensão (localização)



Análise de Dados

O que é um data warehouse?

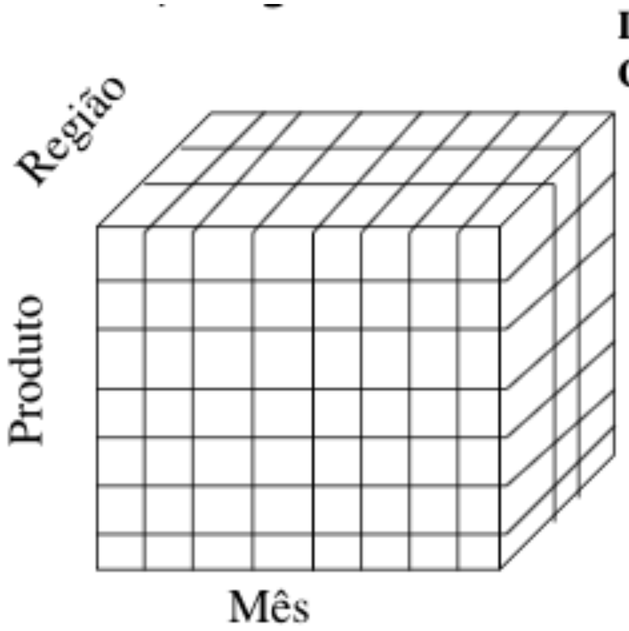
O modelo de dados multi-dimensional

Arquitectura de data warehouses

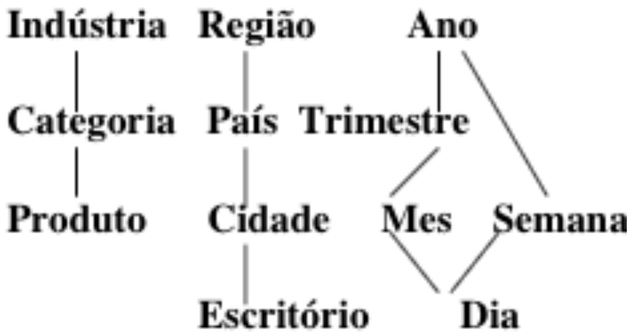
Implementação de data warehouses

Dados Multi-dimensionais

- Volume de vendas como função de produto, mês, região



Dimensões: Produto, Localização, Tempo
Caminhos de sumarização hierárquicos



Análise de Dados

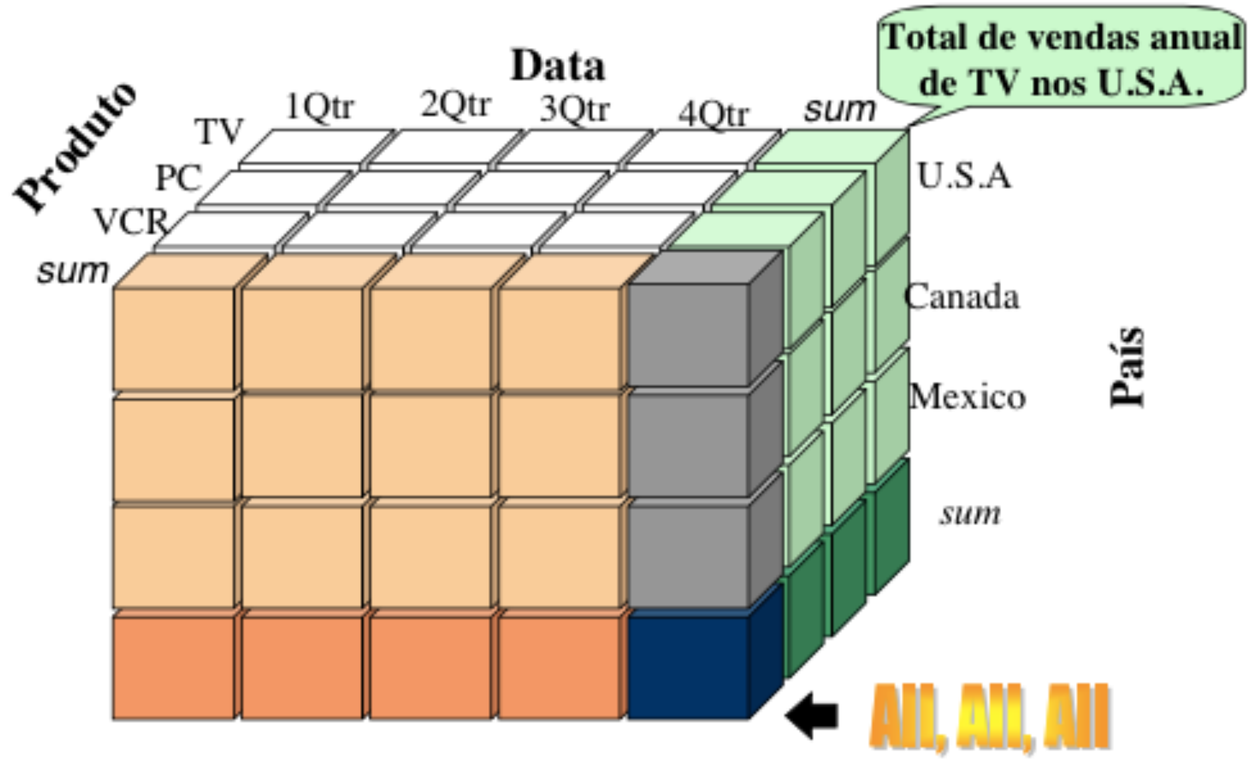
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Exemplo de Cubo de Dados



Análise de Dados

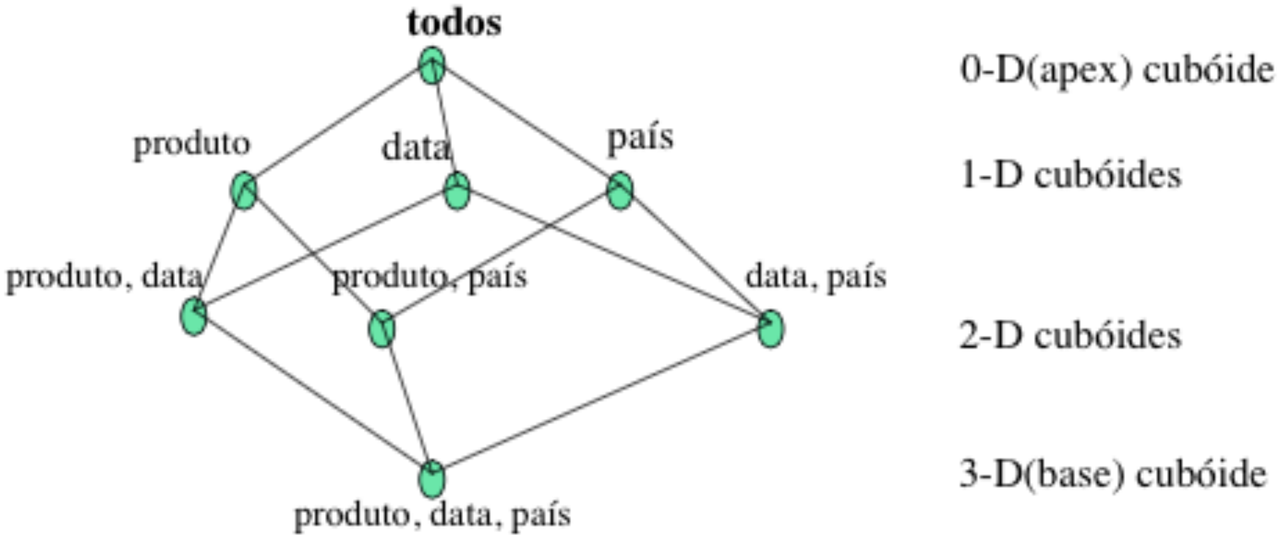
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Cubóides correspondentes ao Cubo



Análise de Dados

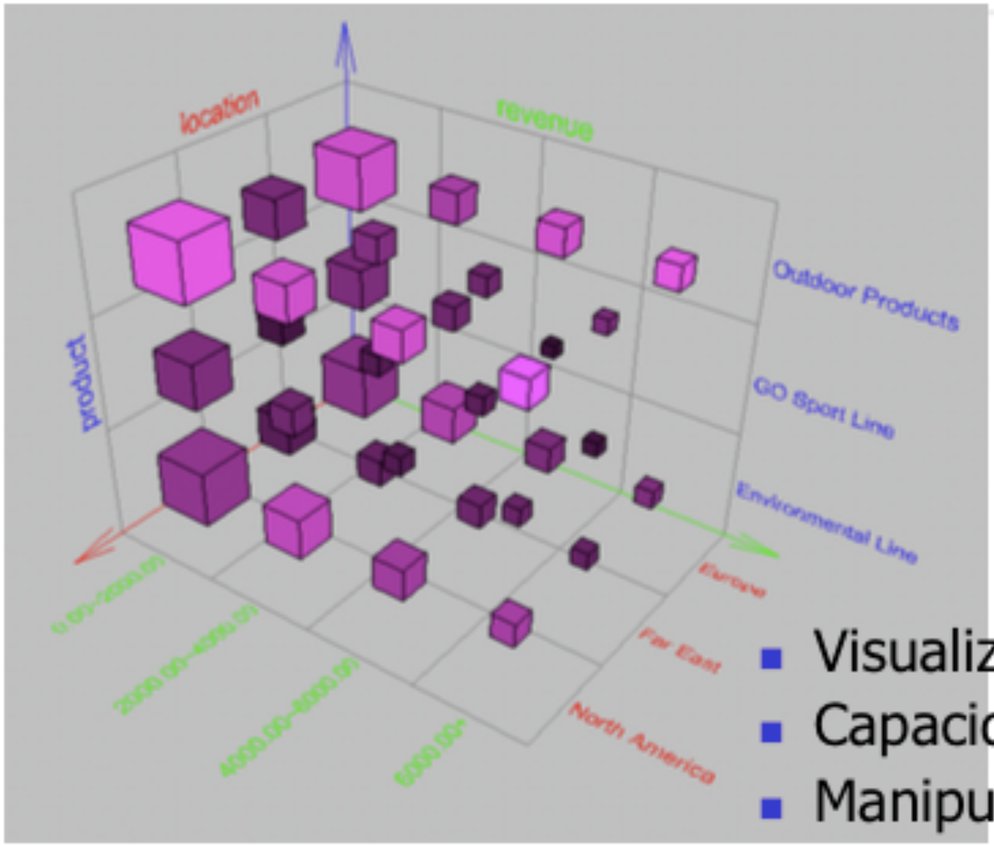
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Pesquisa num Cubo de Dados



- Visualização
- Capacidades OLAP
- Manipulação interactiva

Análise de Dados

Operações OLAP típicas

- **Roll up (drill-up):** sumarizar dados
  - por subida na hierarquia ou por redução de uma dimensão
- **Drill down (roll down)**: inverso de roll-up
  - de sumários de nível mais alto para sumários de nível mais baixo ou mais detalhados, ou pela introdução de dimensões
- **Slice and dice**:
  - project e select
- **Pivot (rotate)**:
  - reorientar o cubo, visualização, de 3D para séries de planos 2D
- Outras operações
  - **drill across:** envolvem mais do que uma tabela de factos
  - **drill through:** do nível mais baixo do cubo para as tabelas relacionais de back-end (usando SQL)

Análise de Dados

Existem variáveis que afectam a escolha do tipo de arquitectura e de implementação de um DW.

- o tempo para a execução do projecto,
- o retorno do investimento,
- a velocidade dos benefícios da utilização da informação,
- a satisfação do utilizador e
- os recursos necessários à implementação de uma arquitectura.

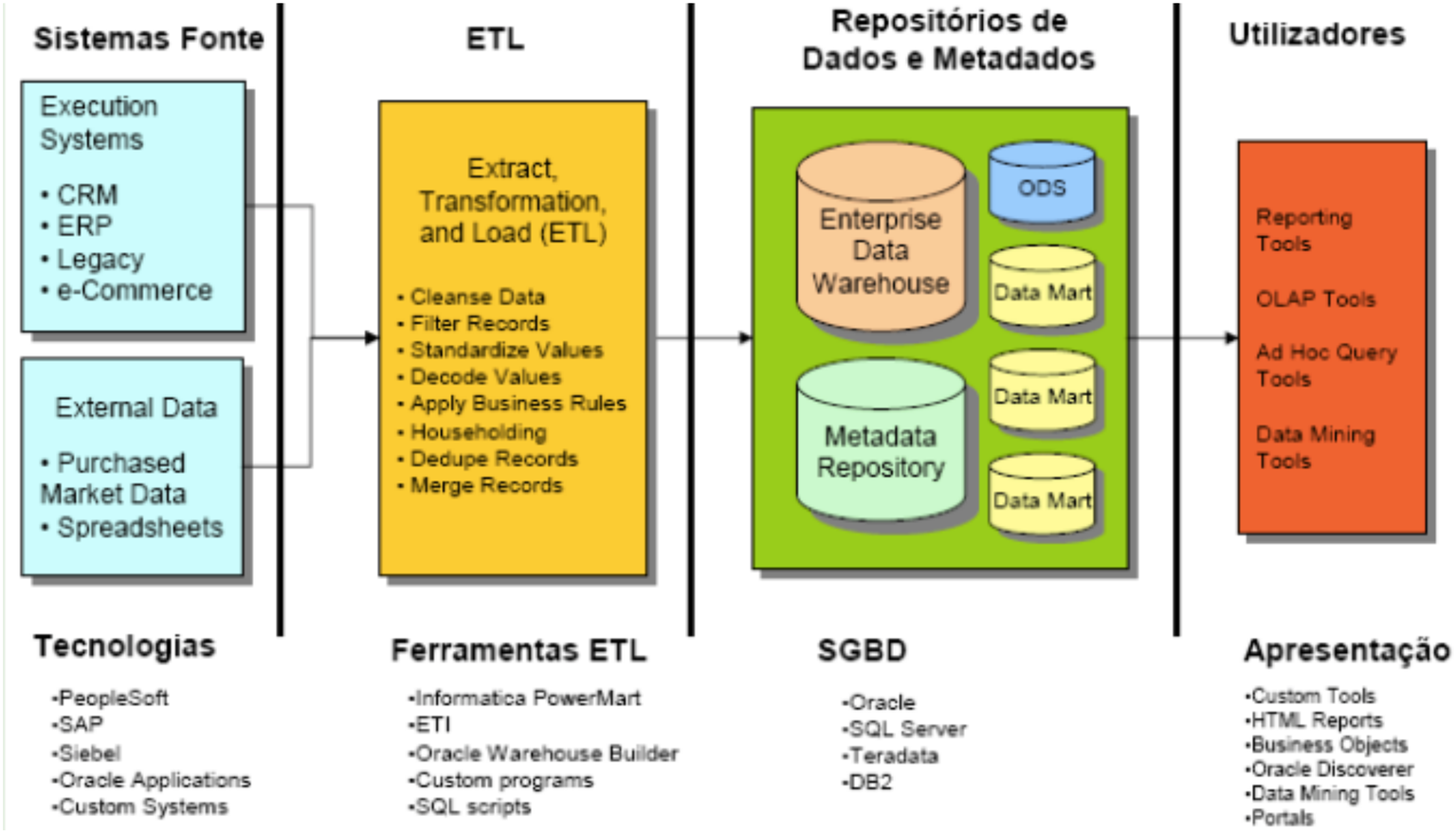
Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses



Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Desenho de Data Warehouses

Quatro perspectivas de desenho de um data warehouse

- **Perspectiva Top-down**
  - Permite a selecção da informação relevante necessária para o data warehouse
- **Perspectiva de Origem de Dados**
  - Mostra a informação a ser adquirida, guardada e gerida por sistema operacionais
- **Perspectiva Data warehouse**
  - consiste em tabelas de factos e tabelas dimensão
- **Perspectiva de Consulta de Análise**
  - vê a perspectiva dos dados no warehouse do ponto de vista do utilizador final

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

- Abordagens Top-down, bottom-up ou uma combinação de ambos
  - **Top-down**: Começa com o desenho e planeamento geral
  - **Bottom-up**: Começa com experiências e protótipos
- Do ponto de vista da engenharia de software
  - **Cascata**: Análise estruturada e sistemática em cada passo antes
  - **Espiral**: Geração rápida e incremental de funcionalidades do sistema
- Processo de desenho típico de data warehouse
  - Escolher um **processo de negócio** a modelar, e.g., encomendas, facturas, etc.
  - Escolher o **grão** (nível de dados atómico) do processo de negócio
  - Escolher as **dimensões** que estão associadas a cada tabela de factos
  - Escolher as **medidas** presentes em cada registo da tabela de factos

Análise de Dados

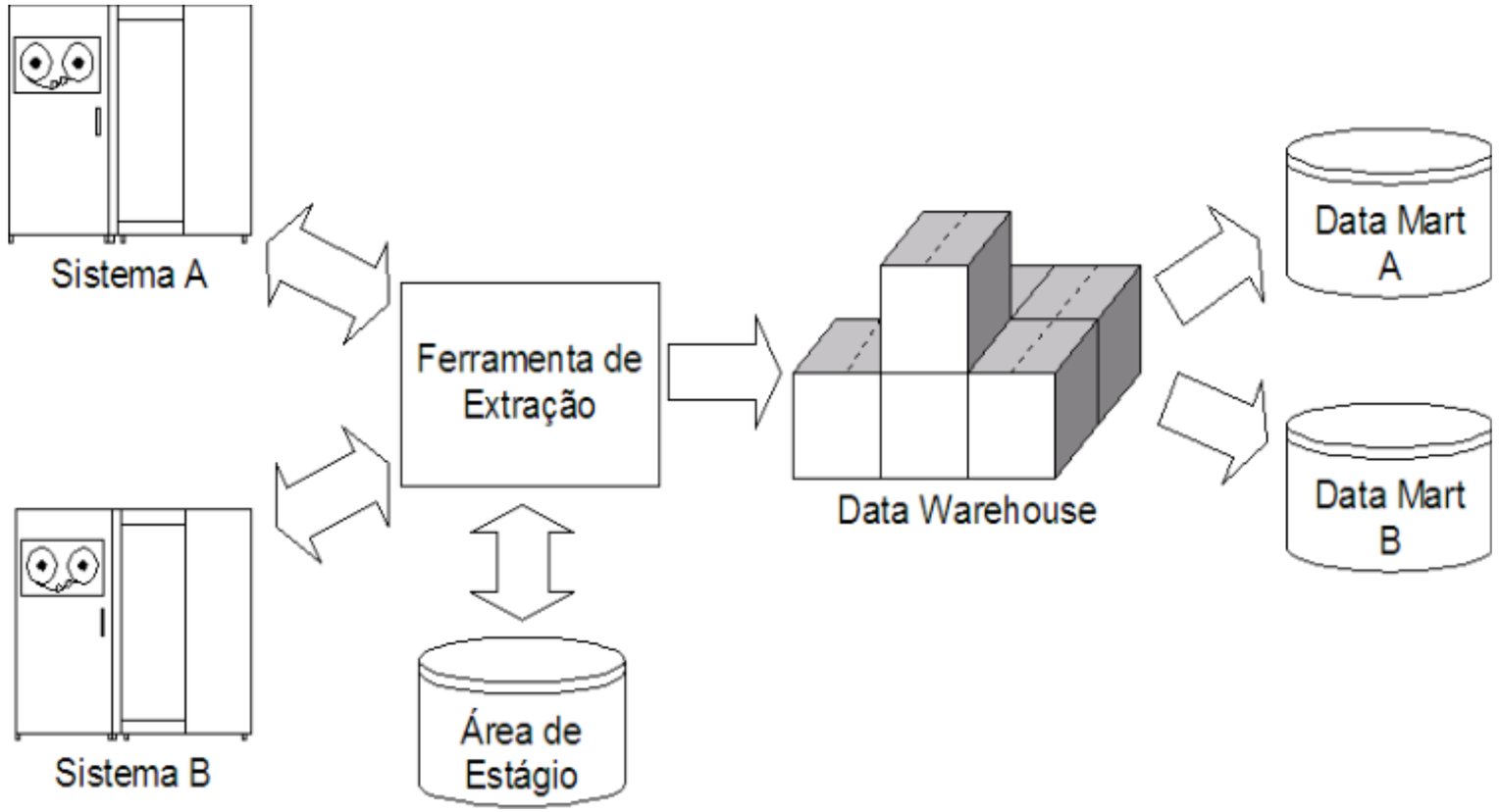
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

- **Top-down (adaptado de Inmon)**



Análise de Dados

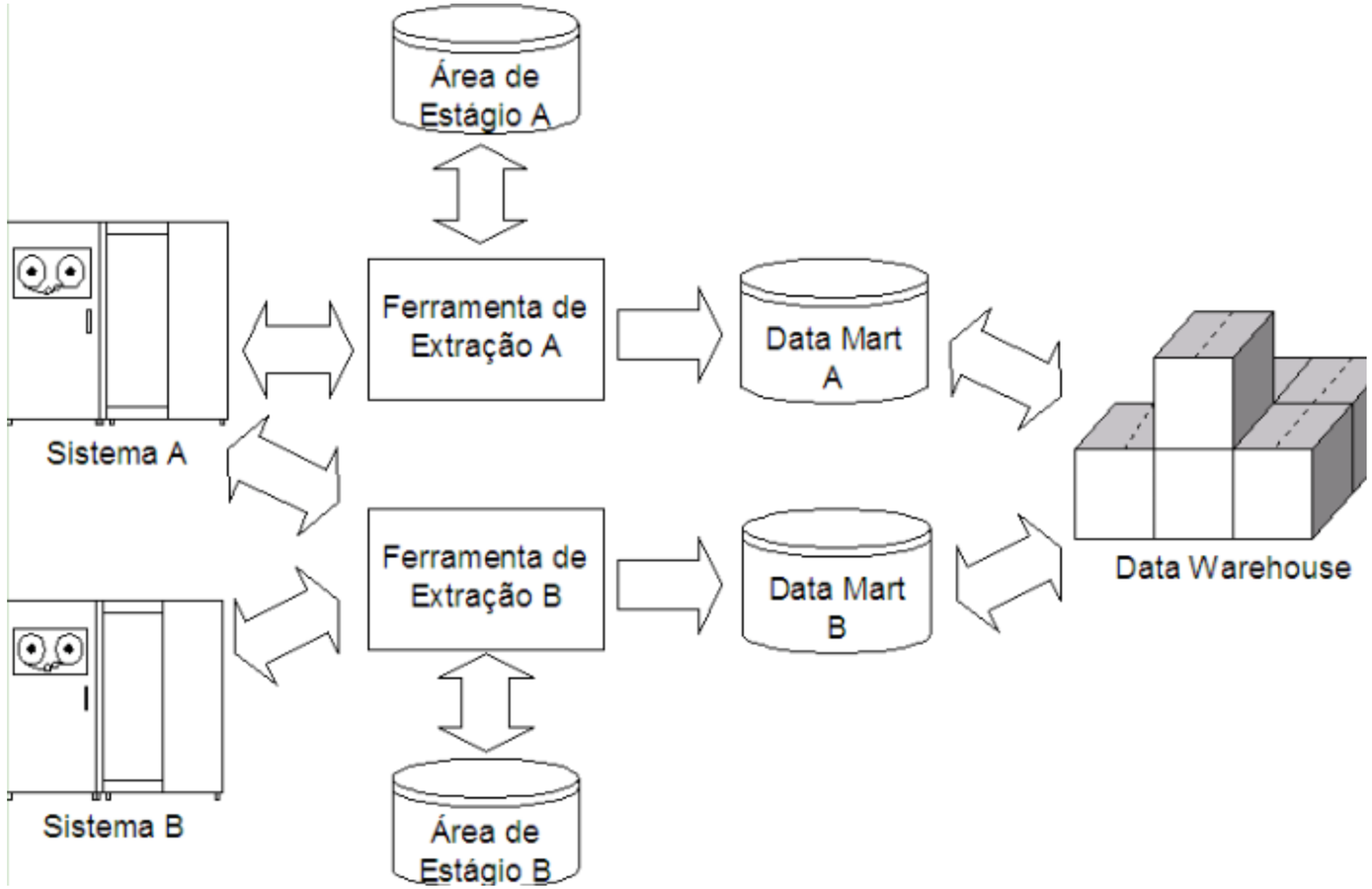
O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

- **Bottom-Up (adaptado de Inmon)**



Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Processo de Modelação

- A modelação dimensional é um processo **top-down**
- Deve adoptar-se a perspectiva do utilizador final (gestor/decisor)
- Escolher as **dimensões** e os **factos** a incluir no DW.

(Kimbal)

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Processo de Modelação

Processo de negócio	É necessário responder à questão: onde e como são recolhidos os dados?
Granularidade	A granularidade determina a dimensionalidade do DW e tem impacto no seu tamanho. Quase sempre faz sentido guardar os dados acerca das dimensões com a maior granularidade possível. O objectivo não é ver cada registo individualmente, mas permitir que as pesquisas sejam mais precisas.
Dimensões da tabela de factos	É, geralmente, possível acrescentar outras dimensões desde que exista apenas um valor dessas dimensões para cada combinação de valores das dimensões já existentes. Se isso não acontecer, é necessário rever a granularidade do modelo.
Medidas na tabela de factos	Na maioria dos casos é uma perda de tempo tentar normalizar as tabelas de dimensão pois iria dificultar a actividade em que o utilizador explora uma única dimensão com o objectivo de definir colunas e restrições para uma pesquisa posterior que consiste em verificar o valor de determinados atributos quando se restringe o valor de outros atributos.

Análise de Dados

- O que é um data warehouse?
- O modelo de dados multi-dimensional
- Arquitectura de data warehouses
- **Implementação de data warehouses**

Processo de Modelação

Dimensão tempo	A dimensão tempo presente em quase todos os DW apresenta vantagens em relação à data em SQL como: permitir análise de dias da semana/fins-de-semana; facilita a divisão em períodos fiscais; permite análise de vendas em feriados e datas especiais; cada registo na tabela representa um dia; a tabela não é construída a partir dos sistemas fonte/operacionais; pode conter dias que ainda não aconteceram e conter campos que permitem análise de dados em períodos temporais.
Hierarquias explícitas	A definição de hierarquias explícitas facilita as operações de *drill-down* e *rollup*, ou seja, em aumentar e diminuir o nível de detalhe de uma consulta.
Tabela factos sem factos	Existem, no entanto, alguns processos de negócio susceptíveis de serem modelados num DW aos quais não existem factos associados. Nestas situações utilizam-se tabelas de factos sem factos. Ou seja usa-se em tabelas de registo de eventos, como por exemplo: para ajudar a perceber que eventos não ocorreram, como por exemplo saber quais os produtos em promoção que não foram vendidos.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Processo de Modelação

Dimensões muito grandes	Este tipo de dimensões necessita de um tratamento diferente de forma a acelerar as pesquisas e a facilitar alterações. Algumas das soluções apontadas passam pela criação de índices apenas nos campos utilizados para fazer pesquisas e na criação de mini dimensões com campos que mudam com frequência. Os campos demográficos são bastante utilizados, quer individualmente, quer em conjunto, para restringir as consultas num DW. A forma mais eficaz de utilizar estes atributos consiste em colocá-los numa mini dimensão separada. Para atributos com carácter contínuo devem utilizar-se gamas de valores e para dimensões muito grandes, as mini dimensões permitem poupar espaço e acelerar as consultas.

Análise de Dados

Os atributos das dimensões podem mudar ao longo do tempo e os registos de uma dimensão podem ter que ser alterados - **Slowly Changing Dimensions (SCD)**. (Kimbal)

- **Sobreposição de valores** - alterar directamente o valor de um ou mais campos errados.
  - vantagem ser fácil de implementar
  - desvantagens a perda dos valores anteriores,
- **Criação de um novo registo** na dimensão
  - generalizar as chaves da dimensão
- **Criação de campos** que permitam registar a evolução dos valores
  - fácil de implementar, os atributos que podem ser modificados;
  - desvantagem, apenas se registam os valores originais e actuais dos campos e os valores intermédios são perdidos.

Análise de Dados

Erros Típicos na Modelação (Kimbal)

- Colocar atributos de texto numa tabela de factos;
- Limitar a descrição dos atributos;
- Dividir hierarquias em múltiplas dimensões;
- Ignorar a necessidade de alterações nas dimensões;
- Resolver os problemas de desempenho à custa de hardware;
- Utilizar chaves operacionais;
- Negligenciar a granularidade da tabela de factos;
- Criar o modelo dimensional com base num relatório; e
- Esperar que os utilizadores pesquisem dados atómicos num formato normalizado.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Erros mais comuns na implantação de um DW

1. Começar o DW com o tipo errado de patrocínio
2. Gerar expectativas que não podem ser satisfeitas
3. Carregar o DW com informações só porque estavam disponíveis
4. Acreditar nas promessas de desempenho dos vendedores de produtos
5. Focalizar dados internos orientados a registo e ignorar o valor de dados externos

Análise de Dados - ETL

O sistema Extract-Transform-Load (ETL) é a base do data warehouse.

Análise de Dados - ETL

Um sistema ETL projetado adequadamente:

- extrai dados dos sistemas de origem,
- reforça a qualidade dos dados e padrões de consistência,
- ajusta dados para que fontes separadas possam ser usadas juntas e
- finalmente entrega dados num formato pronto para apresentação.

Análise de Dados - ETL

Especificamente, o sistema ETL:

- Remove erros e corrige dados perdidos,
- Fornece medidas documentadas de confiança nos dados,
- Captura com segurança o fluxo de dados transacionais,
- Ajusta os dados de várias fontes para serem usados em conjunto, e
- Estruturar os dados para serem conumidos pelas ferramentas do utilizador final.

Análise de Dados - ETL

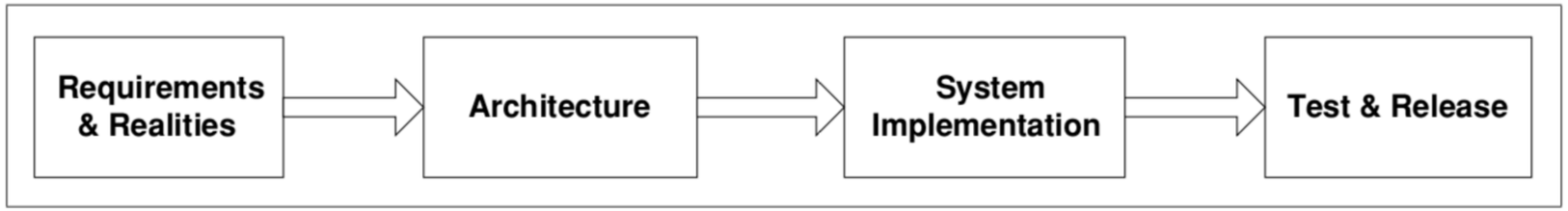
Duas tarefas simultâneas devem ser mantidas ao criar um sistema ETL:

Planeamento e Desenho e

o Fluxo de Dados.

Análise de Dados - ETL

Planeamento e Desenho



Análise de Dados - ETL

O primeiro passo no segmento de Planeamento e Design é a levantamento de todos os requisitos e contextos:

- Necessidades de negócios
- Perfil de dados e outras realidades da fonte de dados
- Requisitos de conformidade
- Requisitos de segurança
- Integração de dados
- Latência de dados
- Arquivo e linhagem

Análise de Dados - ETL

Interfaces de entrega do utilizador final

skills de desenvolvimento disponíveis

skills de gestão disponíveis

Licenças legadas

Análise de Dados - ETL

A segunda etapa desta lista é a da **arquitetura**.

- Hand-coded versus ETL vendor tool
- Batch versus streaming data flow
- Horizontal versus vertical task dependency
- Scheduler automation
- Exception handling
- Quality handling
- Recovery and restart
- Metadata
- Security

Análise de Dados - ETL

A terceira etapa desta lista é a da **implementação**.

- Hardware
- Software
- Coding practices
- Documentation practices
- Specific quality checks

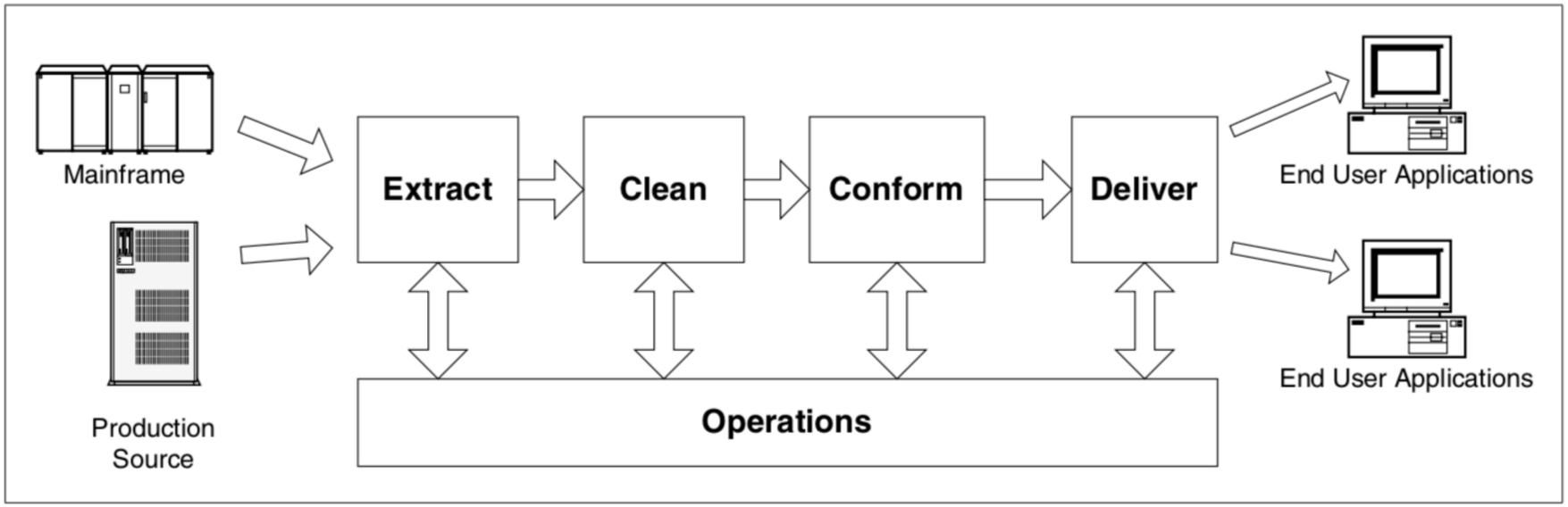
Análise de Dados - ETL

A ultima etapa desta lista, **Testes e Entrega**

- Development systems
- Test systems
- Production systems
- Handoff procedures
- Update propagation approach
- System snapshoting and rollback procedures
- Performance tuning

Análise de Dados - ETL

Fluxo dos Dados



Análise de Dados - ETL

O processo de extração inclui

- Reading source-data models
- Connecting to and accessing data
- Scheduling the source system, intercepting notifications and daemons
- Capturing changed data
- Staging the extracted data to disk

Análise de Dados - ETL

O processo de limpeza contempla

- Reading source-data models
- Connecting to and accessing data
- Scheduling the source system, intercepting notifications and daemons
- Capturing changed data
- Staging the extracted data to disk

Análise de Dados - ETL

O processo de limpeza contempla

- Enforcing column properties
- Enforcing structure
- Enforcing data and value rules
- Enforcing complex business rules
- Building a metadata foundation to describe data quality
- Staging the cleaned data to disk

Análise de Dados - ETL

O processo de conformidade contempla

- Conforming business labels (Tabelas de Dimenão)
- Conforming business metrics and performance indicators (Tabelas de Factos)
- Deduplicating
- Householding
- Internationalizing
- Staging the conformed data to disk

Análise de Dados - ETL

O processo de entrega contempla

- Loading flat and snowflaked dimensions
- Generating time dimensions
  Loading degenerate dimensions
- Loading subdimensions
- Loading types 1, 2, and 3 slowly changing dimensions
- Conforming dimensions and conforming facts
  Handling late-arriving dimensions and late-arriving facts
- Loading multi-valued dimensions
- Loading ragged hierarchy dimensions
- Loading text facts in dimensions
- Running the surrogate key pipeline for fact tables
- Loading three fundamental fact table grains
- Loading and updating aggregations
- Staging the delivered data to disk

Análise de Dados - ETL

O fluxo dos dados básico de quatro etapas é supervisionado pela etapa de **operações**, que se estende desde o início da etapa de **extração** até o final da etapa de **entrega**.

- Scheduling
- Job execution
- Exception handling
- Recovery and restart
- Quality checking
- Release
- Support

Análise de Dados - ETL

The Back Room – Preparing the Data

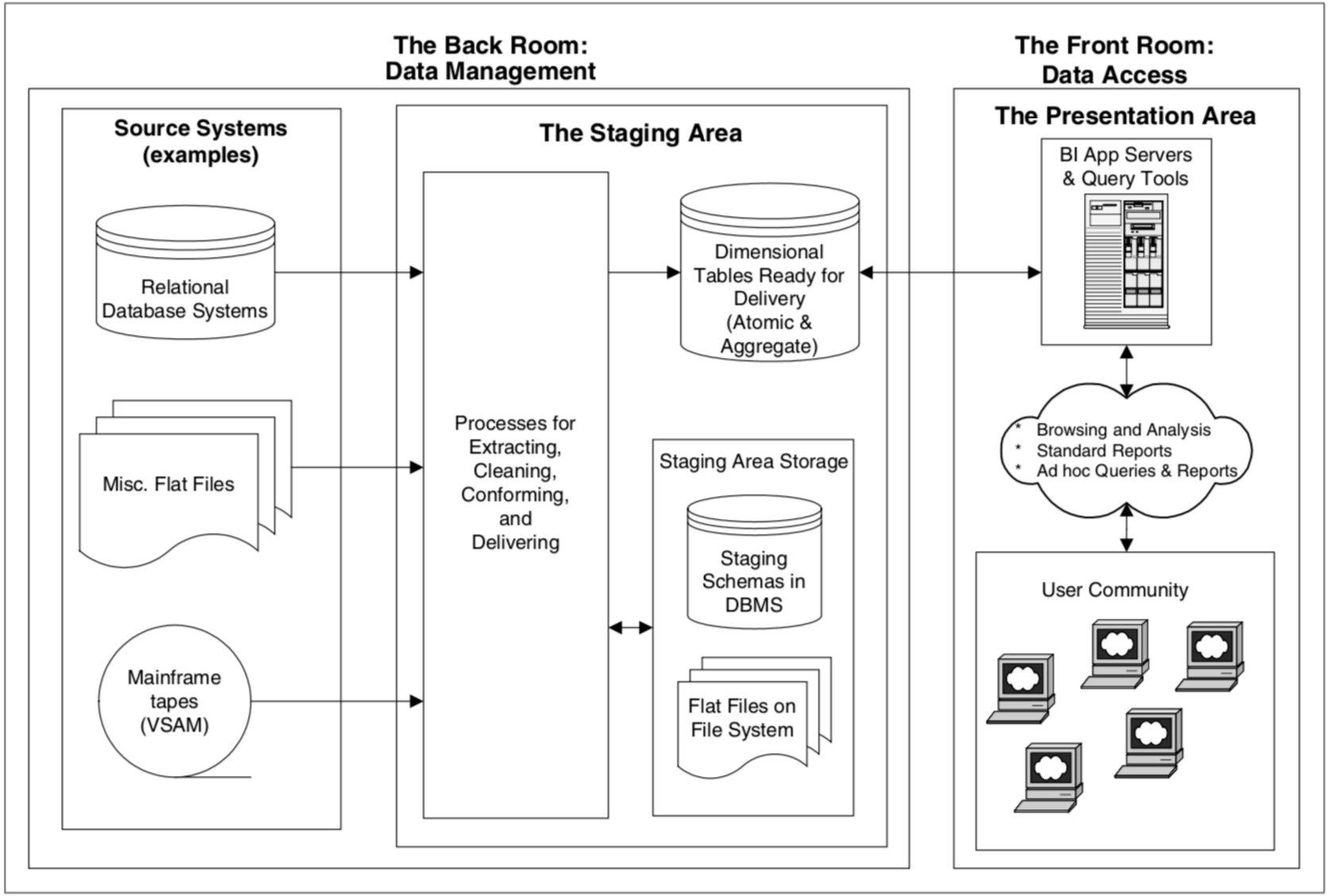


Figure 1.1 The back room and front room of a data warehouse

Análise de Dados

Extração, Transformação e Carregamento de Dados

O processo de ETL dá-se início com a:

- Definição do mapa lógico de dados;
- Descrição dos relacionamentos entre as fontes de dados e os campos destino na DW; e
- Ligação entre o ponto inicial e o ponto final do processo de ETL.

Análise de Dados

Extração, Transformação e Carregamento de Dados

Antes de se implementar o processo de ETL é necessário:

- ter um plano (Mapa Lógico de Dados);
- identificar as fontes de dados candidatas;
- analisar os sistemas fonte;
- percorrer a linhagem dos dados e regras de negócio;
- percorrer o modelo físico de dados; e
- validar cálculos e fórmulas.

Análise de Dados

Extração, Transformação e Carregamento de Dados

O Mapa Lógico de Dados é apresentada da forma:

No destino deve constar:	Na origem deve constar:
- Nome da tabela destino; - Nome da coluna destino; - Tipo de dados da coluna destino; - Tamanho; - Tipo de tabela; - Tabela de Dimensão ou Tabela de Factos; - Tipo de alteração (SCD) – Tipo1, Tipo 2 ou Tipo 3;	- Base de dados origem; - Nome da tabela origem; - Nome da coluna origem; - Tipo de dados da coluna origem.

Na transformação deve constar:
- Descrição exacta da forma como é feita a manipulação dos dados fonte de forma a corresponder ao formato destino que é esperado.

Análise de Dados

Target					Source				Transformation
Table Name	Column Name	Data Type	Table Type	SCD Type	Database Name	Table Name	Column Name	Data Type	
EMPLOYEE_DIM	EMPLOYEE_KEY	NUMBER	Dimension	1				NUMBER	Surrogate key.
EMPLOYEE_DIM	EMPLOYEE_ID	NUMBER	Dimension	1	HR_SYS	EMPLOYEES	EMPLOYEE_ID	NUMBER	Natural Key for employee in HR system
EMPLOYEE_DIM	BIRTH_COUNTRY_NAME	VARCHAR2(75)	Dimension	1	HR_SYS	COUNTRIES	NAME	VARCHAR2(75)	select c.name from employees e, states s, countries c where e.state_id = s.state_id and s.country_id = c.country
EMPLOYEE_DIM	BIRTH_STATE	VARCHAR2(75)	Dimension	1	HR_SYS	STATES	DESCRIPTION	VARCHAR2(255)	select s.description from employees e, states s where e.state_id = s.state_id
EMPLOYEE_DIM	DISPLAY_NAME	VARCHAR2(75)	Dimension	1	HR_SYS	EMPLOYEES	FIRST_NAME	VARCHAR2(75)	select initcap(salutation) || ' '||initcap(first_name) || ' '|| initcap(last_name) from employee
EMPLOYEE_DIM	BIRTH_DATE	DATE	Dimension	1	HR_SYS	EMPLOYEES	DOB	DATE	trunc(DOB)
EMPLOYEE_DIM	SALUTATION	VARCHAR2(12)	Dimension	1	HR_SYS	EMPLOYEES	SALUTATION	VARCHAR2(12)	initcap(salutation)
EMPLOYEE_DIM	FIRST_NAME	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	FIRST_NAME	VARCHAR2(30)	initcap(first_name)
EMPLOYEE_DIM	LAST_NAME	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	LAST_NAME	VARCHAR2(30)	initcap(last_name)
EMPLOYEE_DIM	MARITAL_STATUS	VARCHAR2(12)	Dimension	2	HR_SYS	MARITAL_STATUS	DESCRIPTION	VARCHAR2(12)	select nvl(m.name,'Unknown') from employee e marital_status m where e.marital_status_id = m.marital_status_id
EMPLOYEE_DIM	DIVERSITY_CATEGORY	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	EEO_CLASS	VARCHAR2(30)	decode(eeo_class,null, 'Not Stated', decode(eeo_class,'N', 'Not Stated',eeo_class))
EMPLOYEE_DIM	GENDER	VARCHAR2(12)	Dimension	1	HR_SYS	EMPLOYEES	SEX	VARCHAR2(12)	nvl(sex, 'Unknown')
EMPLOYEE_DIM	EMPLOYEE_STATUS	VARCHAR2(24)	Dimension	1	HR_SYS	EMPLOYEES	STATUS	VARCHAR2(24)	select es.name from employee e employee_status es where e.employee_status_id = m.employee_status_id
EMPLOYEE_DIM	POSITION_CODE	VARCHAR2(12)	Dimension	2	HR_SYS	POSITIONS	POSITION_CODE	VARCHAR2(12)	select p.code from employees e, positions p where p.position_id = e.position_id
EMPLOYEE_DIM	POSITION_CATEGORY	VARCHAR2(30)	Dimension	2	HR_SYS	POSITIONS	POSITION_CATEGORY	VARCHAR2(30)	select p.category from employees e, positions p where p.position_id = e.position_id
EMPLOYEE_DIM	HIRE_DATE	DATE	Dimension	1	HR_SYS	EMPLOYEES	DATE_HIRED	DATE	trunc(date_hired)
EMPLOYEE_DIM	DEPARTMENT_CODE	VARCHAR2(12)	Dimension	2	HR_SYS	DEPARTMENTS	CODE	VARCHAR2(12)	select d.code from employee p, employee_department pd, departments d where p.employee_id= pd.employee_id and pd.department_id = d.department_id
EMPLOYEE_DIM	DEPARTMENT_NAME	VARCHAR2(75)	Dimension	2	HR_SYS	DEPARTMENTS	DESCRIPTION	VARCHAR2(75)	select d.description from employee p, employee_department pd, departments d where p.employee_id= pd.employee_id and pd.department_id = d.department_id
EMPLOYEE_DIM	PART_TIME_FLAG	VARCHAR2(1)	Dimension	1	HR_SYS	EMPLOYEES	PERCENTAGE	VARCHAR2(1)	select decode(sign(percentage-100),-1,'Y','N') from employee
EMPLOYEE_CONTRACT_FACT	EFFECTIVE_DATE_KEY	NUMBER	Fact	N/A	DW_PROD, HR_SYS	DATE_DIM, EMPLOYEE_CONTRACT	DATE_KEY	NUMBER	where employee_contract.eff_date = dw_prod.date_dim.cal_date
EMPLOYEE_CONTRACT_FACT	END_DATE_KEY	NUMBER	Fact	N/A	DW_PROD, HR_SYS	DATE_DIM, EMPLOYEE_CONTRACT	DATE_KEY	NUMBER	where employee_contract.end_date = dw_prod.date_dim.cal_date
EMPLOYEE_CONTRACT_FACT	CURRENCY_KEY	NUMBER	Fact	N/A	DW_PROD, HR_SYS	CURRENCY_DIM, EMPLOYEE_CONTRACT	CURRENCY_KEY	NUMBER	where employee_contract.currency_code = dw_prod.currency_dim.currency_code
EMPLOYEE_CONTRACT_FACT	RATE_TYPE_KEY	NUMBER	Fact	N/A	DW_PROD, HR_SYS	RATE_TYPE_DIM, EMPLOYEE_CONTRACT	RATE_TYPE_KEY	NUMBER	where employee_contract.rate_type_id = dw_prod.rate_type_dim.rate_type_id
EMPLOYEE_CONTRACT_FACT	PROJECT_KEY	NUMBER	Fact	N/A	DW_PROD, HR_SYS	PROJECT_DIM, EMPLOYEE_CONTRACT	PROJECT_KEY	NUMBER	where employee_contrac.project_code = dw_prod.project_dim.project_code
EMPLOYEE_CONTRACT_FACT	EMPLOYEE_ROLE_KEY	NUMBER	Fact	N/A	DW_PROD, HR_SYS	EMPLOYEE_ROLE_DIM, EMPLOYEE_CONTRACT	EMPLOYEE_ROLE_KEY	NUMBER	where employee_contract.employee_role = employee_role_dim.contractor_role_name
EMPLOYEE_CONTRACT_FACT	CONTRACT_TYPE_KEY	NUMBER	Fact	N/A	DW_PROD, HR_SYS	CONTRACT_TYPE_DIM, EMPLOYEE_CONTRACT	CONTRACT_TYPE_KEY, CONTRACT_TYPE_ID	NUMBER	where dw_prod.employee_contract.contract_type = contract_type_dim.contract_type_id
EMPLOYEE_CONTRACT_FACT	CONTRACT_NUMBER	NUMBER	Fact	N/A	DW_PROD, HR_SYS	EMPLOYEE_CONTRACT	CONTRACT_NUMBER	NUMBER	Degenerate Dimension
EMPLOYEE_CONTRACT_FACT	RATE_AMOUNT_LOCAL	NUMBER	Fact	N/A	DW_PROD, HR_SYS	EMPLOYEE_CONTRACT	AMOUNT	NUMBER	sum(amount)
EMPLOYEE_CONTRACT_FACT	RATE_AMOUNT_USD	NUMBER	Fact	N/A	DW_PROD, HR_SYS	EMPLOYEE_CONTRACT	AMOUNT	NUMBER	select ec.amount * avg(cc.conversion_rate) from employee_contract ec, currency_conversion cc where ec.currency = cc.from_currency and cc.effective_Date between ec.effective_date and ec.end_date group by ec.amount

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

PROCESSO DE EXTRACÇÃO DE DADOS

Consiste no processo de

- compreender,
- seleccionar e
- copiar

os dados fonte para a DAS (área de tratamento dos dados)

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Deteção de anomalias nos dados (exemplo):

- Valores nulos em chaves estrangeiras e valores nulos noutras colunas;
- datas em campos que não representam datas, pois existem vários formatos para as datas.
- A extracção é efectuada a partir de diferentes plataformas, sendo necessário a integração de dados de fontes heterogéneas.
- As dimensões e factos conformes garantem a coesão da DW.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Capturar as modificações nos dados fonte é crucial. (Kimbal)

- Timestamps (triggers)
  - É feita a adição de uma coluna na qual é registada a hora/data da alteração de cada registo.
- Partições
  - As tabelas de dados são divididas em partições e cada partição representa um horizonte temporal.
- Processo de eliminação (Área de retenção)
  - Carregamento inicial e incrementais
    Criam-se duas tabelas na DAS (previous_load e current_load).

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

PROCESSO DE TRANSFORMAÇÃO DE DADOS

- limpeza dos dados;
- eliminação de campos inúteis;
- combinação de dados provenientes de fontes diferentes;
- criação de chaves primárias independentes dos sistemas fonte e
- construção de agregados de modo a acelerar as pesquisas.

A limpeza e conformidade geram metadados (tabelas de erros), metadados acompanham os dados até estes chegarem aos utilizadores finais do DW.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

O objectivo:

- qualidade dos dados,
- onde os dados devem ser:
- correctos (os valores dos dados são genuínos),
- claros (os dados só podem ter um significado),
- consistentes (utilizar apenas uma convenção para a representação dos dados) e,
- terem completude (os valores dos campos existirem).

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

- **Limpar dados "sujos"**
  - valores sem sentido,
  - ausência de dados,
  - dados duplicados,
  - dados cujo significado não é claro (e que os metadados não esclarecem),
  - dados contraditórios e
  - dados que violam regras de integridade.
- **2. Eliminar inconsistências**
  - insuficiências no processo de extracção,
  - alterações nos sistemas operacionais e
  - problemas técnicos nos sistemas operacionais.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Tipos de transformações a dois níveis.

- 1. Ao nível do registo:
  - **selecção**: particionamento dos dados;
  - **junção**: combinação dos dados e
  - **agregação**: resumo dos dados.

- 2. Ao nível dos campos:
  - envolvendo um único campo: de um campo para outro campo e,
  - envolvendo múltiplos campos: de muitos campos para um ou de um campo para muitos.

Análise de Dados

PROCESSO DE CARREGAMENTO DE DADOS

- Geralmente são carregados muitos registos de uma só vez (bulk loading) e
- depois de carregados os dados são indexados.
- O carregamento das tabelas são uma fase crítica em que eventuais falhas podem levar a recuperações complexas.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

PROCESSO DE CARREGAMENTO DE DADOS

Tudo o que é feito para optimizar o desempenho da DW tende a atrasar o carregamento, como por exemplo:

- índices,
- agregados,
- particionamento de tabelas,
- paralelismo e
- distribuição.

Análise de Dados

PROCESSO DE CARREGAMENTO DE DADOS

- O carregamento inicial permite a disponibilização na DW dos dados extraídos das fontes operacionais e correctamente validados na DAS.
- Para além do carregamento inicial é necessário resolver os carregamentos periódicos, com características diferentes, podendo ser usados na actualizações de dimensões.
- Será necessário a considerar: duração estimada do carregamento e o impacto na coerência da DW caso o processo tenha de ser interrompido.

Análise de Dados

PROCESSO DE CARREGAMENTO DE DADOS

Passos típicos

- Planeamento
  - definir o mapa lógico de dados,
  - a infra-estrutura para a área de estágio,
  - escolher as ferramentas de ETL
- Carregamento de dimensões
  - Dimensões estáticas e simples
  - Dimensões que mudam e
  - tratar todos os restantes casos como dimensões geradas com dados manuais.
- Carregamento de factos
- Elaborar e testar processo de carregamentos periódicos.
- Automatizar o processo ao máximo
- Utilizar ferramentas sofisticadas de suporte.

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

Implementação de data warehouses

Infra-estrutura para Área de Estágio

- A infra-estrutura para a DAS pode ir de uma simples conta no servidor onde vai ficar a DW,
- A máquinas dedicadas de grande capacidade.

Análise de Dados

Tarefas de administração:

- construir, utilizar e manter as ferramentas de extracção de dados dos sistemas operacionais,
- garantir a qualidade dos dados (após cada extracção),
- construir e manter agregados e
- vigiar e afinar o desempenho do sistema. ´

Exemplos:

- fazer cópias de segurança periodicamente e
- recuperar o estado da base de dados em caso de falha,
- construir e manter templates para exploração de dados,
- formar e treinar utilizadores.

Análise de Dados

Problemas

No processo de extracção:
- acesso aos sistemas fonte,
- identificação dos novos dados a carregar,
- minimizar a interferência nos sistemas fonte e
- lidar com as alterações nos sistemas fonte.

No processo de transformação:
- definição de regras para as transformações de dados,
- limpeza de dados para automatizar e
- métricas de qualidade de dados.

No processo de carregamento:
- minimizar a janela de carregamento. (Kimbal)

Análise de Dados

O que é um data warehouse?

O modelo de dados multi-dimensional

Arquitectura de data warehouses

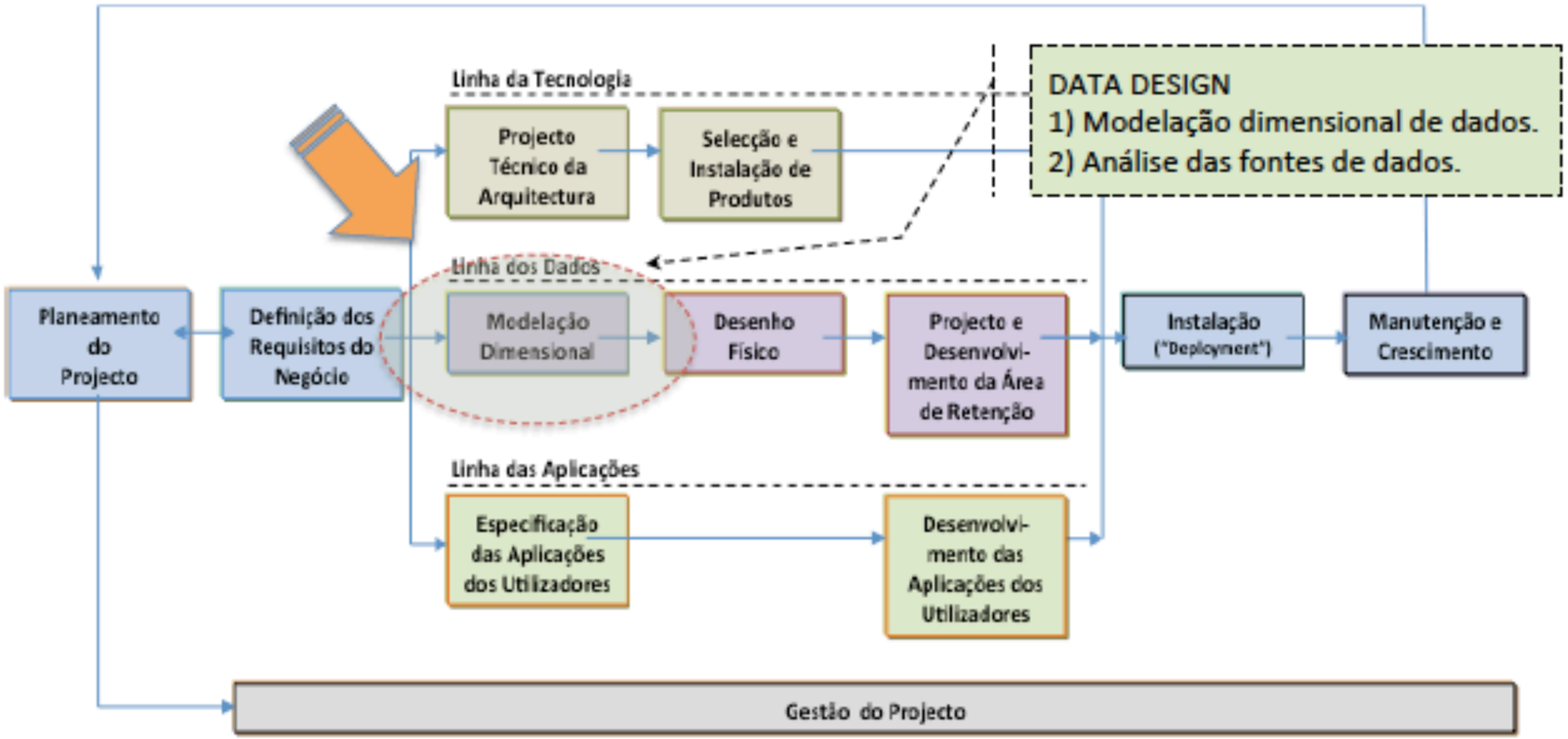
Implementação de data warehouses

TAREFAS DE ADMINISTRAÇÃO E GESTÃO (mais importantes)

- Monitorizar carregamentos de dados de várias fontes;
- Verificar a qualidade e integridade dos dados;
- Gerir e actualizar os metadados;
- Monitorizar o desempenho de modo a garantir tempos de resposta às pesquisas aceitáveis;
- Garantir uma eficiente utilização de recursos;
- Realizar a auditoria e relatórios acerca da utilização do DW;
- Replicar e distribuir os dados;
- Manter uma gestão eficiente do DW;
- Limpar os dados;
- Arquivar e fazer cópias de segurança;
- Implementar mecanismos de recuperação de falhas;
- Gerir a segurança e as prioridades;
- Gerir o espaço de armazenamento do DW;
- Efectuar estatísticas;
- Criar índices;
- Criar agregados;
- Reconstruir on-line os índices;
- Criar partições;
- Providenciar o hardware.

Análise de Dados

O Ciclo de Desenvolvimento



(Kimball, et al., 1998)

Análise de Dados

A Modelação Dimensional

1. Construção da matriz de decisão
2. Seleção do *datamart* a desenvolver.
3. Escolha do grão das tabelas de factos.
4. Escolha das dimensões de análise.
5. Desenvolver o diagrama das tabelas de factos.
6. Documentar as tabelas de factos.
7. Projetar o detalhe das dimensões.
8. Desenvolver os diversos factos derivados.
9. Revisão do projeto com os utilizadores e sua aceitação.
10. Revisão das recomendações de ferramentas *enduser* para o projeto da base de dados.

Análise de Dados

A Modelação Dimensional

11. Revisão das recomendações de sistemas de gestão de bases de dados para o projeto da base de dados.
12. Completar o esquema lógico da base de dados.
13. Identificar os possíveis candidatos de agregados armazenados previamente.
14. Desenvolver a estratégia de desenvolvimento para as tabelas de agregados.
15. Revisão do esquema lógico da base de dados.
16. Certificar o esquema desenvolvido para a base de dados com o fornecedor das ferramentas para suporte à decisão.
17. Rever o projeto e tratar da aceitação por parte dos utilizadores.

Análise de Dados

Análise das fontes

1. Identificar as fontes de dados candidatas.
2. Analisar o conteúdo das fontes de dados – dados e metadados.
3. Desenvolver uma tabela com o mapeamento dos dados entre as diversas fontes de dados operacionais e os dados do *data warehouse – source-to-target-map*.
4. Estimar o número de registos envolvidos futuramente no processo de povoamento.
5. Rever o projeto e tratar da aceitação por parte dos seus futuros utilizadores.

Análise de Dados

O Processo de Modelação

O processo de modelação dimensional que iremos seguir acompanhar. de perto a abordagem proposta por ***Kimball e Ross (2002)***, desenvolvendo, passo a passo, requisito a requisito, de forma a desenvolver um esquema multidimensional cobrindo todos os tipos de objetos de dados – **tabelas de facto, dimensões, tabelas ponte e medidas** – que podemos encontrar neste tipo de esquemas.

Análise de Dados

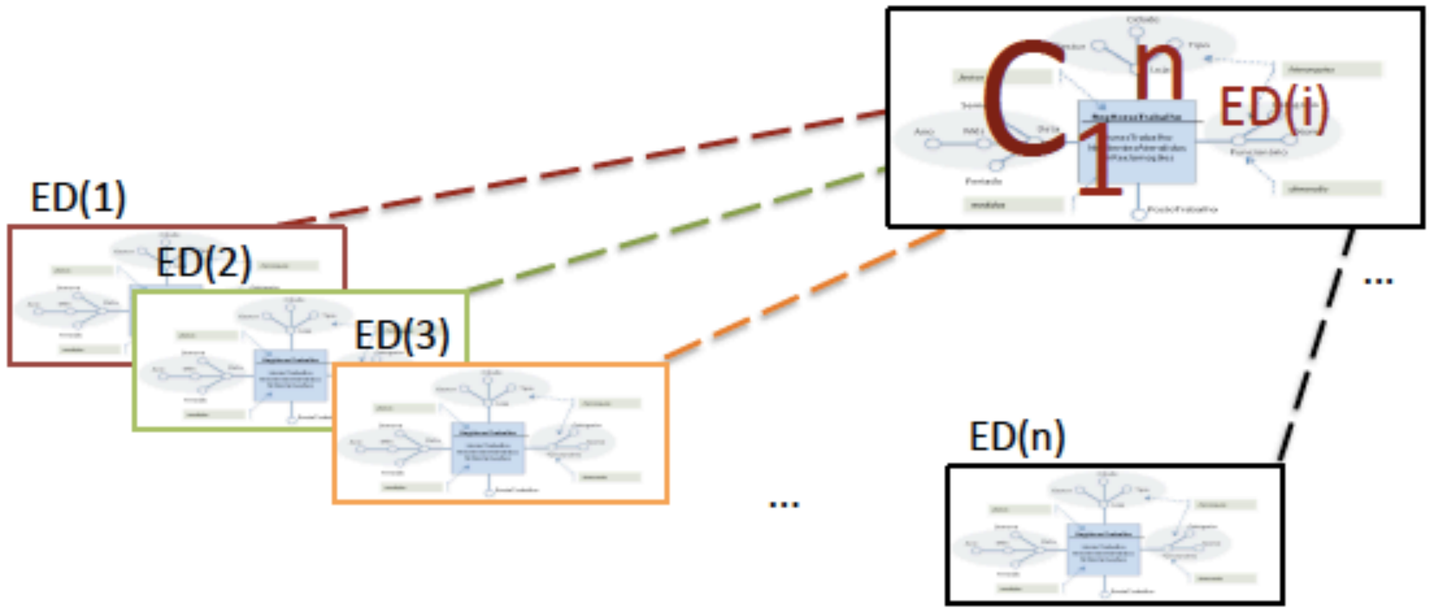
Um Desenvolvimento ***Bottom-Up***

Uma das formas mais usuais de fazer o desenvolvimento de um esquema dimensional é através da utilização do método dos “**4 passos**” (Kimball e Ross 2002), que pressupõe o desenvolvimento do sistema de ***data warehousing*** tipicamente de baixo para cima (***bottom-up***).

Os “nossos” ***data warehouse*** serão projetados área a área, sendo desenvolvido de forma incremental e tomando em consideração todos os (sub)esquemas dimensionais desenvolvidos até ao momento.

Análise de Dados

Bottom-Up



Análise de Dados

O Método do “**4 Passos**”

Os quatro passos do método (Kimball e Ross, 2002) (Imhoff, et al, 2003):

1. Seleção da área de suporte à decisão a implementar.
2. Definição do detalhe dos factos (o grão) do processo selecionado.
3. Seleção das dimensões de análise sobre as quais se pretende analisar os factos.
4. Definição das medidas a integrar na estrutura de cada facto.

Análise de Dados

O Método do “**4 Passos**”

Os quatro passos do método (Kimball e Ross, 2002) (Imhoff, et al, 2003):

1. Seleção da área de suporte à decisão a implementar.
2. Definição do detalhe dos factos (o grão) do processo selecionado.
3. Seleção das dimensões de análise sobre as quais se pretende analisar os factos.
4. Definição das medidas a integrar na estrutura de cada facto.

Análise de Dados

Referências

- Kimbal, J.; The Data Warehouse Lifecycle Toolkit; John Wiley & Sons; 2002.
- Kimbal, J.; The Data Warehouse ETL Toolkit; John Wiley & Sons; 2004.
- Inmon; Building the Data Warehouse; 3rd Ed. W. C. Publishing, John Wiley & Sons; 2002.
- Machado, Felipe; Tecnologia e Projecto de Data Warehouse, Érica; 2007.