Conceitos Básicos

Profa. Dra. Cristina Dutra de Aguiar Ciferri

Data Warehousing

Engloba arquiteturas, algoritmos e ferramentas que possibilitam que dados selecionados de provedores de informação autônomos, heterogêneos e distribuídos sejam integrados em uma única base de dados, conhecida como data warehouse (DW)

Data Warehouse

- Coração do ambiente de data warehousing
- Banco de dados
  - voltado para o suporte aos processos de gerência e tomada de decisão
  - tem como principais objetivos prover eficiência e flexibilidade na obtenção de informações estratégicas e manter os dados sobre o negócio com alta qualidade

Características dos Dados

- Orientados a assunto
  - relativos aos temas de negócio de maior interesse da corporação
  - *exemplos*: clientes, produtos, promoções, contas e vendas
- Integrados
  - dados obtidos dos provedores de informação corrigidos para eliminar possíveis inconsistências

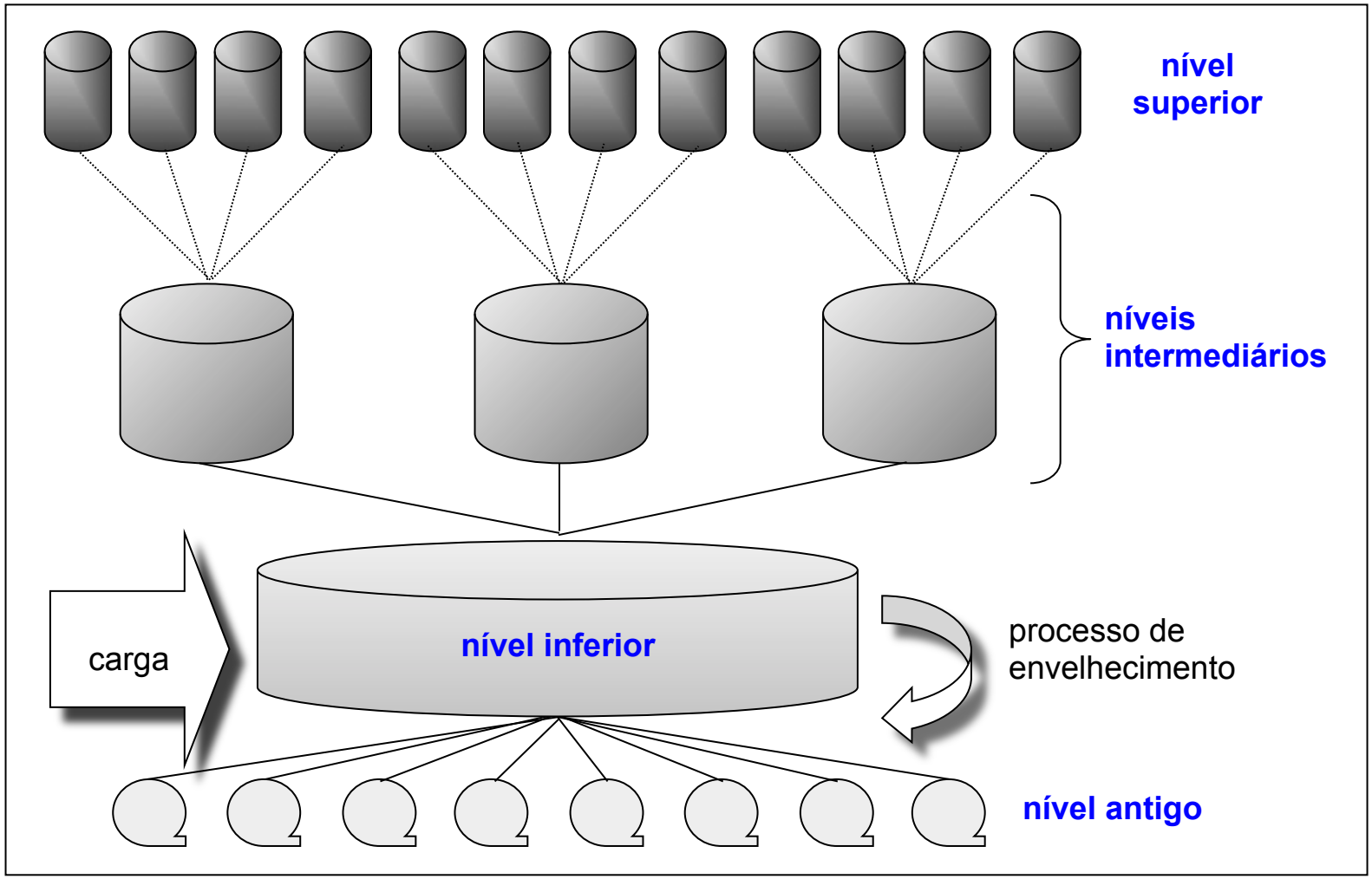
Características dos Dados

- Não voláteis
  - o conteúdo do DW permanece estável por longos períodos de tempo
- Históricos
  - relevantes a algum período de tempo
  - *exemplo*: usualmente dados relativos a um grande espectro de tempo (5 a 10 anos) encontram-se disponíveis

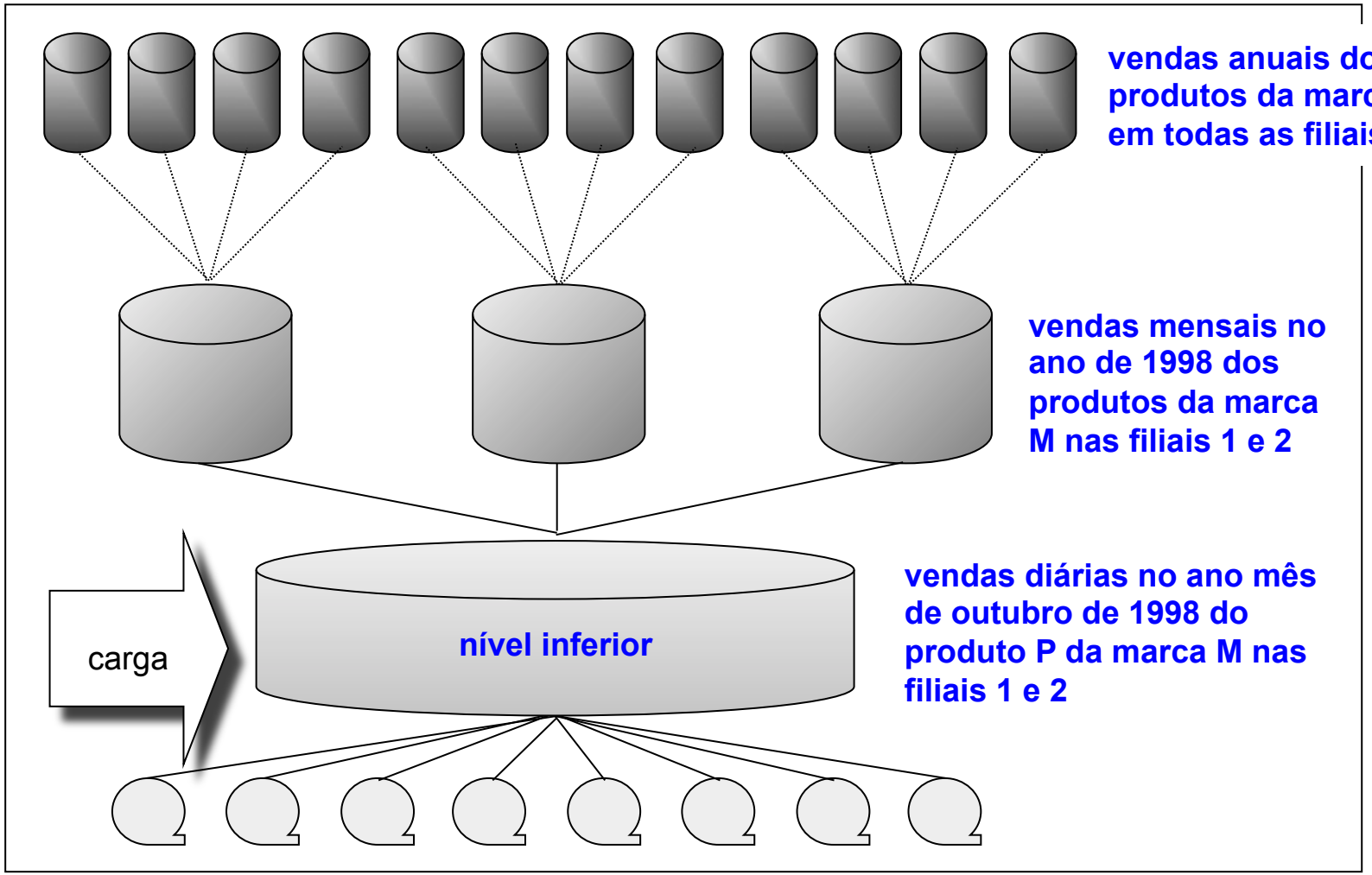
Características dos Dados

- Organizados em diferentes níveis de agregação
  - nível inferior: dados primitivos coletados do ambiente operacional
  - níveis intermediários: dados com graus de agregação crescente
  - nível superior: dados altamente resumidos (agregados)

Níveis de Agregação



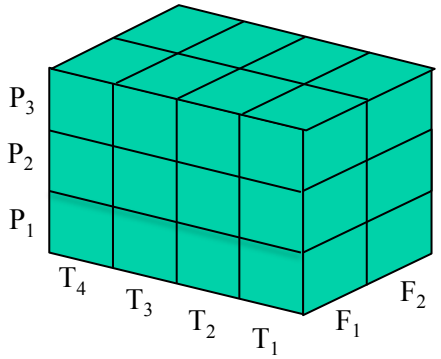
Níveis de Agregação



Modelagem Multidimensional

- Análises dos usuários de SSD
  - representam requisições multidimensionais aos dados do DW
    - visualização dos dados segundo diferentes perspectivas
  - permitem a identificação de problemas e de tendências
- Exemplo
  - vendas por *produto* por *filial* por *tempo*

(Hiper)cubo de Dados Multidimensional (Vendas)



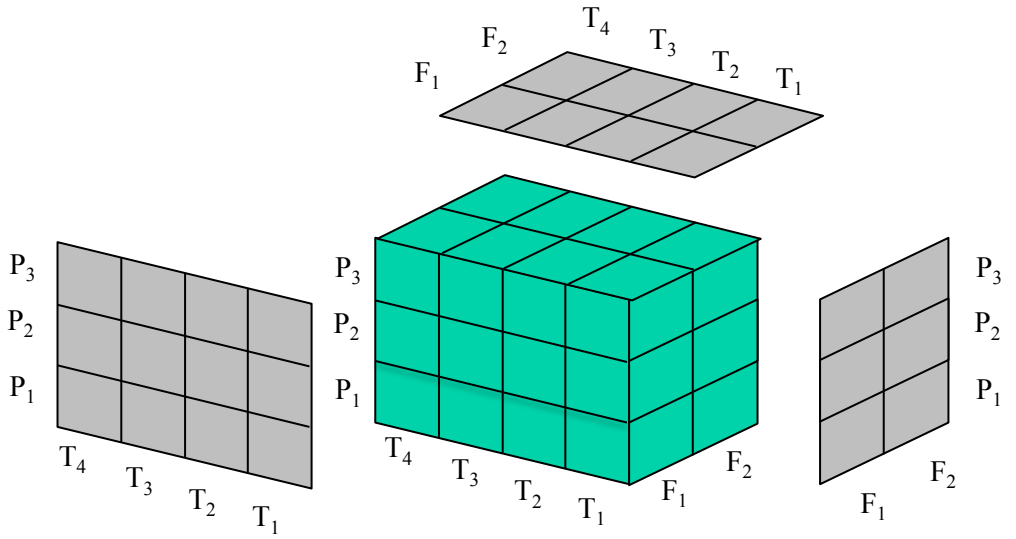
visão multidimensional *vendas* por produto por filial por tempo

- representação gráfica
- semântica subjacente

(Hiper)cubo de Dados Multidimensional (QualidadeAr)

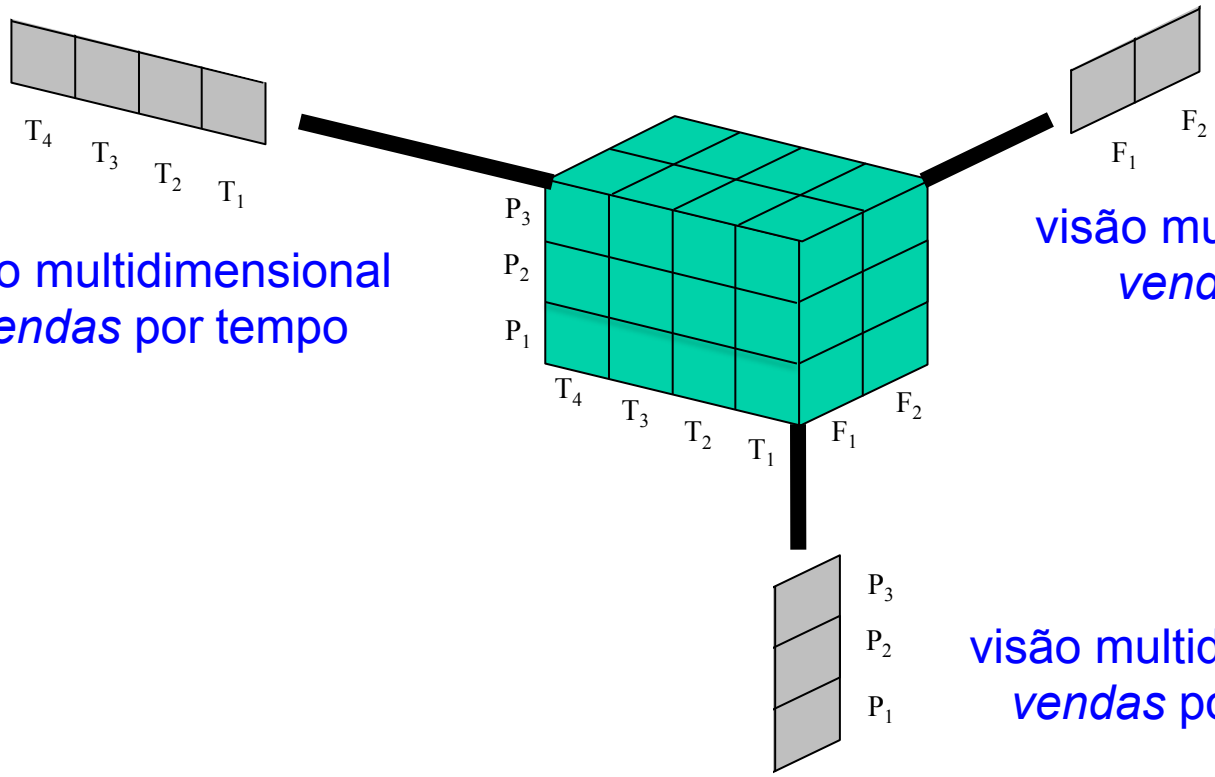
visão multidimensional *vendas* por tempo por filial

visão multidimensional *vendas* por produto por tempo



visão multidimensional *vendas* por produto por filial

(Hiper)cubo de Dados Multidimensional (QualidadeAr)

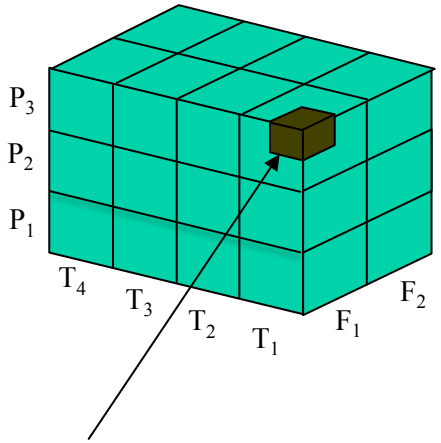


visão multidimensional *vendas* por tempo

visão multidimensional *vendas* por filial

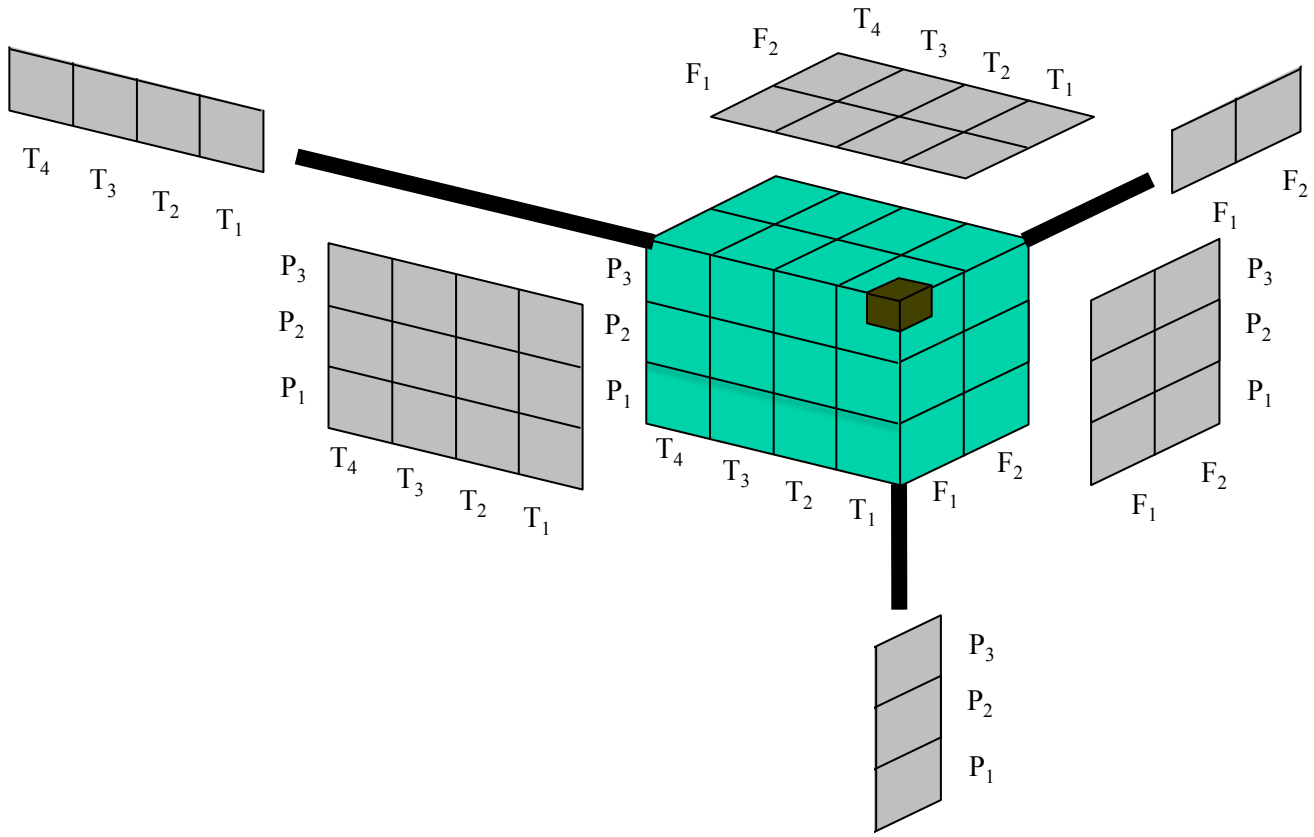
visão multidimensional *vendas* por produto

(Hiper)cubo de Dados Multidimensional (QualidadeAr)

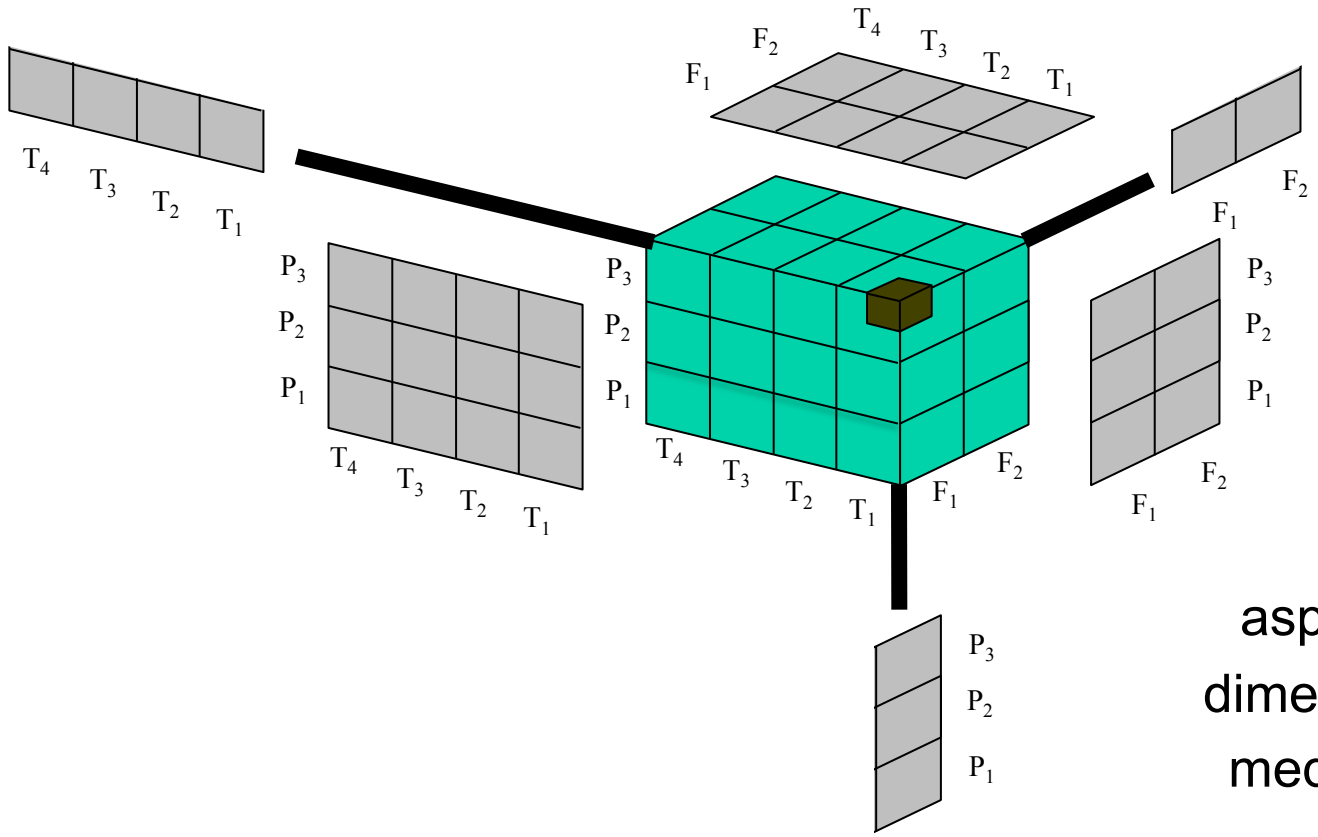


visão multidimensional
vendas

(Hiper)cubo de Dados Multidimensional (QualidadeAr)



(Hiper)cubo de Dados Multidimensional (QualidadeAr)



aspectos estáticos

dimensões (atributos)

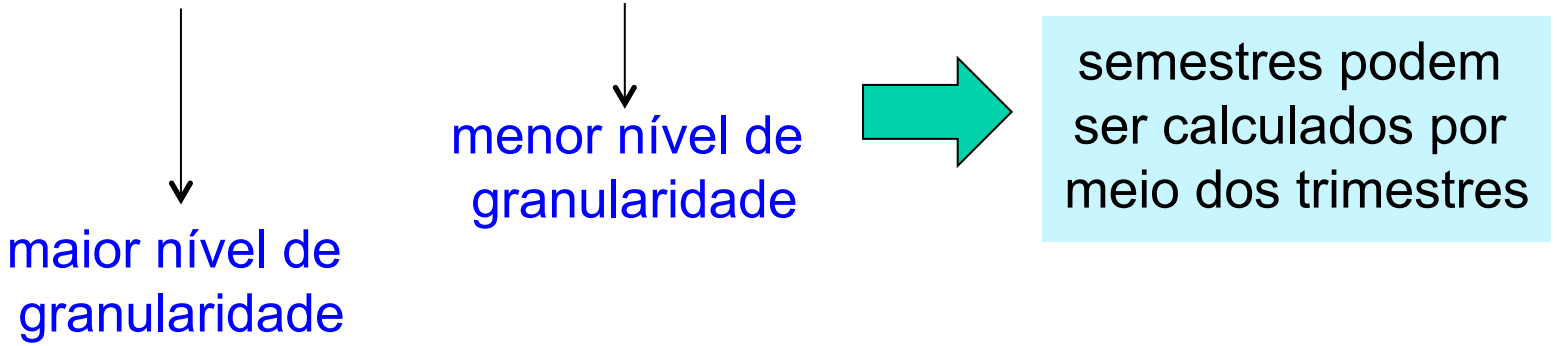
medidas numéricas

Dimensão

- Representa uma perspectiva de análise dos usuários de SSD
- Composta por atributos
- *Exemplo*: dimensão tempo
  - *atributos*: dia, mês, trimestre, semestre, ano
  - *semântica*: dia 15/09/2015, do mês de setembro, do terceiro trimestre, do segundo semestre, de 2015.

Ordenação Parcial ($\preceq$)

- Significado
  - um atributo de maior nível de granularidade de uma hierarquia de atributos pode ser determinado usando um atributo de menor nível de granularidade dessa hierarquia
- *Exemplo*: dimensão tempo
  - (semestre) $\preceq$ (trimestre)



Exemplos

- Dimensão tempo
  - *atributos*: dia, mês, trimestre, semestre, ano
  - (ano) $\preceq$ (semestre) $\preceq$ (trimestre) $\preceq$ (mês) $\preceq$ (dia)
  - *atributos*: dia, mês, quadrimestre, ano
  - (ano) $\preceq$ (quadrimestre) $\preceq$ (mês) $\preceq$ (dia)

podem ser definidas uma ou mais hierarquias de atributos por dimensão

Definição Formal

- Lattice (Reticulado de Cuboides) <L, $\preceq$>, onde
  - L: conjunto de visões (ou agregações)
  - $\preceq$: relação de dependência
- Propriedades de L
  - contém pelo menos a visão do nível inferior
  - pode conter uma visão completamente agregada (*all/none/vazio*), a qual pode ser calculada a partir de qualquer outra visão

Harinarayan, V. Rajaraman, A., Ullman, J. D. Implementing Data Cubes Efficiently. In Proceedings of the ACM SIGMOD International Conference of Management of Data, p. 205-216,1996.

Definição Formal

- Para três visões v, w, u, podem ser definidas as seguintes funções:

$$\text{ancestrais}(v) = \{\, w \mid v \preceq w \,\} .$$

$$\text{descendentes}(v) = \{\, w \mid w \preceq v \,\} .$$

$$\text{ancestrais\_diretos}(v) = \text{pais}(v) = \{\, w \mid v \prec w, \nexists\, u, v \prec u, u \prec w \,\} .$$

$$\text{descendentes\_diretos}(v) = \text{filhos}(v) = \{\, w \mid w \prec v, \nexists\, u, w \prec u, u \prec v \,\} .$$

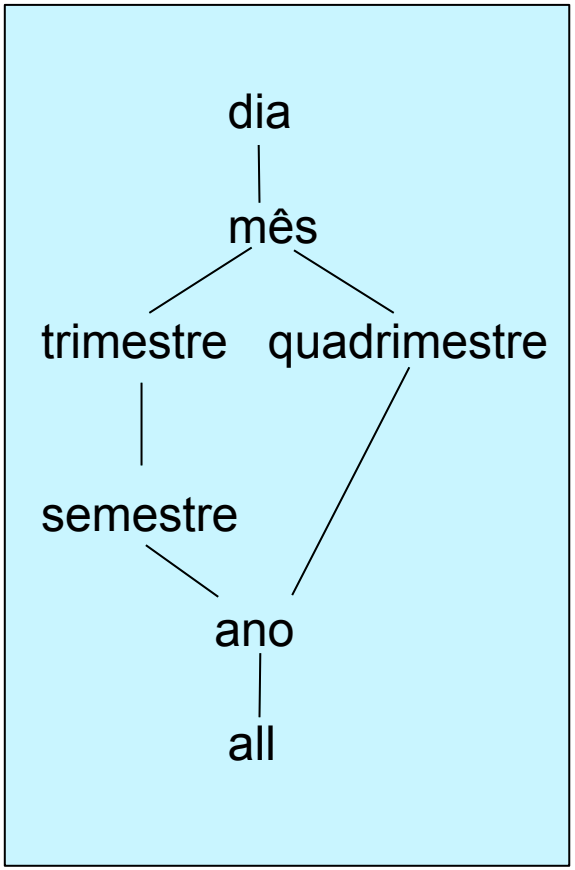
sendo que $v \prec w \implies v \preceq w \wedge v \neq w$.

Exemplo

- Dimensão tempo

(all) $\preceq$ (ano) $\preceq$ (semestre) $\preceq$ (trimestre) $\preceq$ (mês) $\preceq$ (dia)

(all) $\preceq$ (ano) $\preceq$ (quadrimestre) $\preceq$ (mês) $\preceq$ (dia)

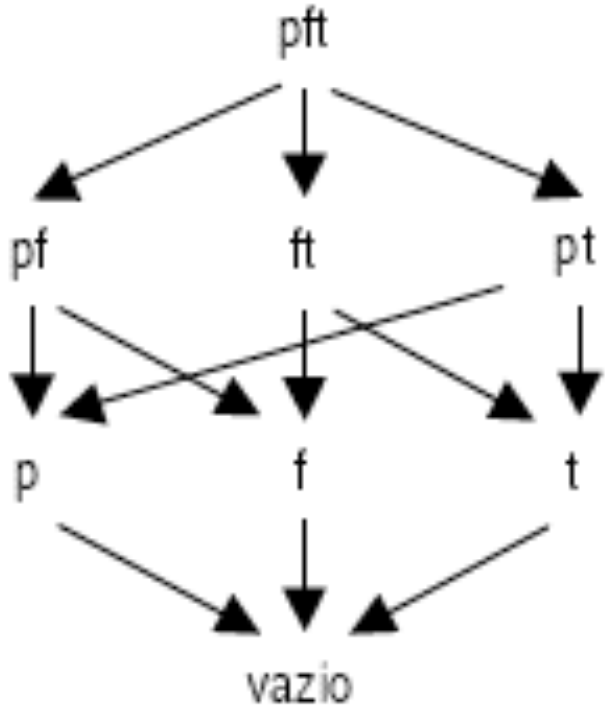


níveis de agregação
inferior
superior

Grafo de Derivação

- Grafo orientado
  - par (V,E) de conjuntos disjuntos de vértices V e arestas E
  - mapeamentos
    - inic: E → V e term: E → V
      - cada aresta *e* sai de um vértice inicial inic(*e*) e chega a um vértice terminal term(*e*)
      - *e* é direcionada de inic(*e*) para term(*e*).
  - Característica
    - não possui ciclos nem arestas múltiplas
- Usado para representar o lattice

Exemplo



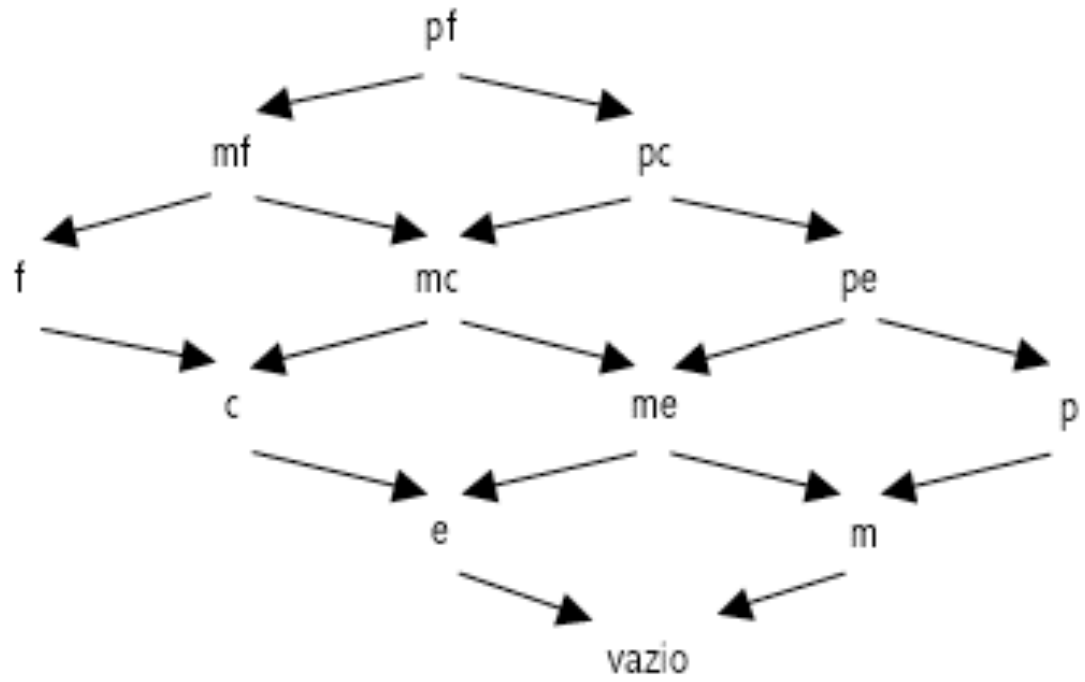
níveis de agregação

inferior

superior

dimensões produto (p), filial (f), tempo (t)

Exemplo



níveis de agregação

inferior

superior

dimensões produto (p), filial (f)

hierarquias de atributos: all $\preceq$ marca (m) $\preceq$ p

all $\preceq$ estado (e) $\preceq$ cidade (c) $\preceq$ f