# SAD – Dados, Datawarehouses, e OLAP

**4ºAno M, AN,FZ,EN-MEC,EN-AEL**

V 2.0, V.Lobo, EN/ISEGI, 2021

## Tipos de dados e operações básicas

1

## Dados numéricos

- Inteiros ou reais
- Precisão e gama dinâmica
  - Número de bits
  - Tipo de representação
    - Vírgula fixa, vírgula flutuante, números astronómicos
- Operações
  - Relações de ordem, operações aritméticas
- Exemplos
  - Temperaturas, nº de pessoas, etc
  - 34, 24.5, $20.4 \times 10^{-15}$, 32144152353, …
- Dados numéricos multidimensionais
  - Vectores numéricos

2

## Dados numéricos

- Como comparar vectores numéricos ?
  - Distâncias d(***x,y***)
    - 3 condições formais:
      - $d(x,y) \geq 0, \forall x,y, \; e \; d(x,y) = 0, \Rightarrow x=y$
      - $d(x,y) = d(y,x), \forall x,y$
      - $d(x,y) \leq d(x,z) + d(z,y), \forall x,y,z$
    - Exemplos
      - Distância Euclideana (dimensão *n*)

$$d(\boldsymbol{x},\boldsymbol{y}) = \left( \sum_{j=1}^{n} (x_j - y_j)^2 \right)^{1/2}$$

3

## Distâncias entre vectores

- Distâncias de Minkowski de **ordem *p***

$$d(x,y) = \left( \sum_i |x_i - y_i|^p \right)^{1/p}$$

  - *Ordem 1 – Distância de manhatten, ou "city block"*

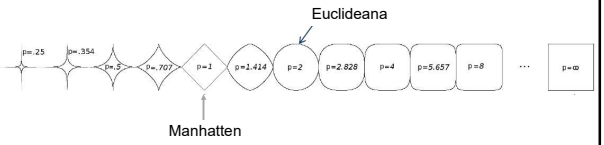
$$d(x,y) = \sum |x_i - y_i|$$

  - *Ordem 2 – Distância Eulideana*
  - *Ordens mais altas*
    - *Dependem cada vez mais da componente mais diferente*
    - *Úteis para evitar "outliers"*

4

## Distâncias entre vectores

- Qual a região que está a uma distância de 1 de um dado ponto, usando diferentes índices *p* nas distâncias de Minkowsky num espaço bi-dimensional ?




5

## Distâncias entre vectores

- Distâncias ponderadas
  - Dão pesos diferentes a componentes diferentes

$$d(\mathbf{x},\mathbf{y}) = \left( \sum \varphi_i (x_i - y_i)^p \right)^{1/p}$$

  - Se o factor de ponderação fôr a matriz de correlação e a ordem fôr 2, teremos a distância de Mahalanobis, ou distância euclideana normalizada

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{y}-\mathbf{x})}$$

ou simplificando:

$$d(x,y) = \left( \sum_i \frac{|x_i - y_i|^p}{\sigma^2} \right)^{1/p}$$

- Produto interno (semelhança em vez de distância)
  - São uma medida de correlação entre os vectores
  - São a projecção de um vector sobre o outro

$$d(\boldsymbol{x},\boldsymbol{y}) = \sum x_j y_j = \boldsymbol{x}|\boldsymbol{y} = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos\theta$$

6

# SAD – Dados, Datawarehouses, e OLAP

**4ºAno M, AN,FZ,EN-MEC,EN-AEL**

V 2.0, V.Lobo, EN/ISEGI, 2021

## Distâncias entre vectores

- Máxima correlação

$$d(x,y) = \max_k \sum x_i y_{i-k}$$

- Cosenos directores
  - É sensível à relações entre as componentes e não à sua magnitude

$$d(x,y) = \cos\theta = \frac{\sum x_i y_i}{\|x\| \times \|y\|}$$

- Outras
  - Menor diferença
  - Maior diferença
  - Tanimoto (aplicado a reais)

$$d(x,y) = \frac{\sum x_i y_i}{\|x\|^2 + \|y\|^2 - \sum x_i y_i}$$

7

## Dados categóricos

- Booleanos
  - Só têm valor 0 ou 1
  - Exemplos
    - Tem a altura mínima, tem um curso, tem...
- Ordinais
  - Têm um número finito de valores
  - Os valores têm uma relação de ordem (mas não podem ser feitas operações aritméticas)
  - Exemplos
    - Escalões de vencimentos, Escalas de comportamento
    - Mau/Suficiente/Bom/Muito Bom, Alto/médio/baixo…
- Categóricos (puros)
  - Não têm relação de ordem
  - Exemplos
    - Naipes de cartas, raças,
    - Paus/Ouros/Espadas/Copas, Marinha/Administração Naval/Fuzileiros/…

8

## Distâncias entre vectores categóricos

- Distância de Hamming
  - Número de bits diferentes
  - Equivalente à distância de manhatten ou ao quadrado da distância euclideana
  - Exemplo
    - D(0010, 1010)=1, D(0010,1101)=4
- Distância de edição ou de Levenshtein
  - Número de alterações (apagar um valor ou acrescentar um valor)
  - Exemplo
    - D(ABC,AB)=1, D(ABC,AD)=3

9

## Distâncias entre vectores categóricos

- Tabela de contingência entre valores dos vectores

| | | Object x | | |
|---|---|---|---|---|
| | | 1 | 0 | sum |
| Object y | 1 | ***a*** | ***b*** | a+b |
| | 0 | ***c*** | ***d*** | c+d |
| | sum | a+c | b+d | a+b+c+d |

- Métricas:

| Coefficients | Equation | Range |
|---|---|---|
| Simple Matching (Sokal and Michener 1958) | $\frac{a+d}{a+b+c+d}$ | [0,1] |
| Russel and Rao (Russel and Rao 1940) | $\frac{a}{a+b+c+d}$ | [0,1] |
| Rogers and Tanimoto (Rogers and Tanimoto 1960) | $\frac{a+d}{a+d+2(b+c)}$ | [0,1] |
| Hamann (Hamann 1961) | $\frac{(a+d)-(b+c)}{a+b+c+d}$ | [-1,1] |
| Ochiai II (Ochiai 1957) | $\frac{ad}{\sqrt{(a+c)(a+b)(d+c)(d+b)}}$ | [0,1] |
| Sokal and Sneath (Sokal and Sneath 1963) | $\frac{2(a+d)}{2(a+d)+b+c}$ | [0,1] |

| Coefficients | Equation | Range |
|---|---|---|
| Jaccard (Jaccard 1901) | $\frac{a}{a+b+c}$ | [0,1] |
| Anderberg (Anderberg 1973) | $\frac{a}{a+2(b+c)}$ | [0,1] |
| Czekanowsky / Sorensen-Dice (Dice 1945) | $\frac{2a}{2a+b+c}$ | [0,1] |
| Kulczynski I (Kulczynski 1927) | $\frac{a}{b+c}$ | [0,+∞] |
| Kulczynski II (Kulczynski 1927) | $\frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right)$ | [0,1] |
| Ochiai (Ochiai 1957) | $\frac{a}{\sqrt{(a+b)(a+c)}}$ | [0,1] |

10

## Medidas de semelhança/dissemelhança

- Não obedecem às 3 condições das distâncias
  - Podem não ser simétricas
  - Podem ser o inverso de uma distância
  - Podem não respeitar a desigualdade triangular
- Exemplos
  - Algumas das métricas do acetato anterior
  - "Distância" de Kullback–Leibler

$$d(x,y) = \sum x_i \log \frac{x_i}{y_i}$$

11

## Outros tipos de dados

- Conjuntos
  - Podem ser semelhantes a dados categóricos
    - Representados e manipulados como categóricos
  - Podem ser conjuntos de pontos
    - Representados como listas
    - Distância de Hausdorff
      - Maior das menores distâncias de um conjunto ao outro

$$d(x,y) = \max(\min_j d(x_i, y_j))$$

- Árvores ou outros grafos
- Mapas
- Etc,etc,etc…

12

# SAD – Dados, Datawarehouses, e OLAP

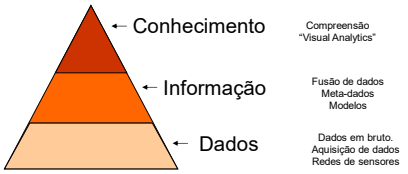
**4ºAno M, AN,FZ,EN-MEC,EN-AEL**
V 2.0, V.Lobo, EN/ISEGI, 2021

## Organização dos dados

13

## Informação é poder...

- "Água é vida"...
  - Todos os anos morre gente afogada...
- É necessário "trabalhar" a informação
- Hierarquia de compreensão e utilidade




14

## SI Operacional *vs* Analítico

- Sistema de Informação **Operacional**
  - Ligado directamente aos **processos**
  - Processamento em tempo real, **contínuo**
  - Muitos dados, **pouco processamento**
  - Constante **mutação**
  - Dia a dia da operação
- Sistema de Informação **Analítico**
  - Ligado aos **decisores**
  - Processamento "off-line", em **tempo diferido**
  - Muitos dados e **MUITO processamento**
  - Maior **estabilidade**
  - Memória da organização

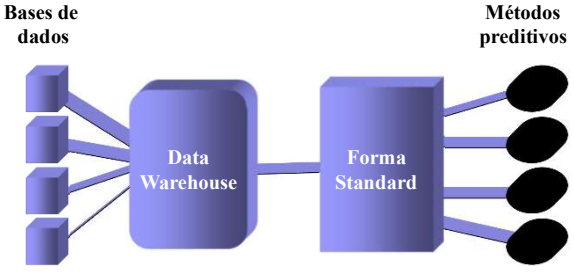
15

## Datawarehouse

- Definição de W.H.Inmon
  - *A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.*

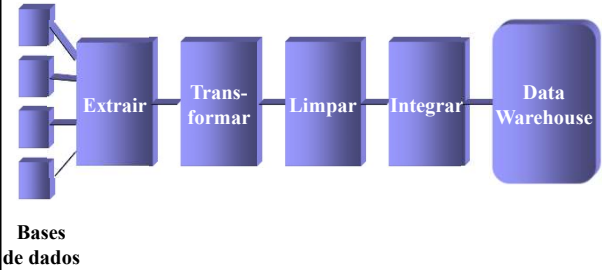
16

## O modelo de "data warehouse"




17

## Passos para construir a "data warehouse" (processo de ETL)



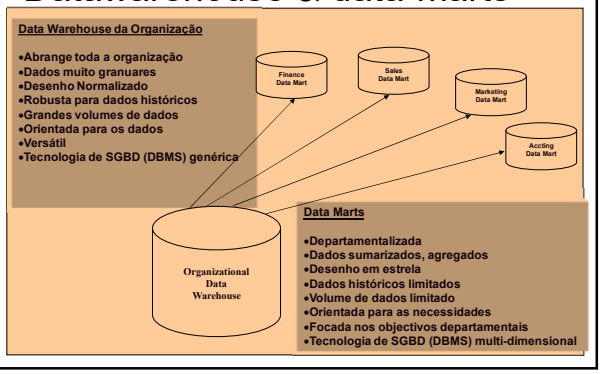

18

# SAD – Dados, Datawarehouses, e OLAP

**4ºAno M, AN,FZ,EN-MEC,EN-AEL**

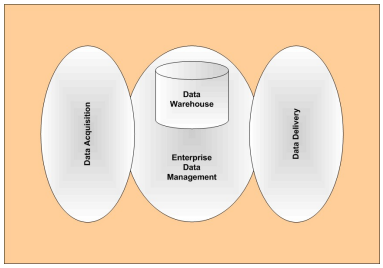
V 2.0, V.Lobo, EN/ISEGI, 2021

## Datawarehouse & data-marts




19

## Outras perspectivas….




20

## Medição, indicadores, visualização

- Relatórios "tradicionais"
  - Relatórios contabilísticos, tabelas de resultados

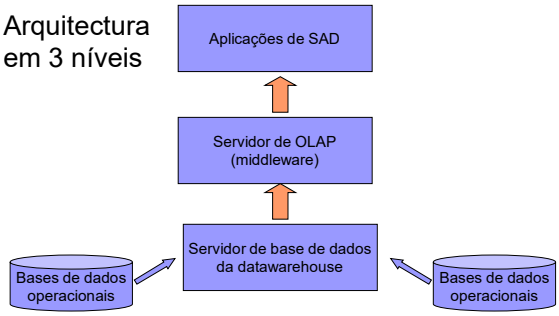
- **Dashboards**
  - Conceito de "tableau de bord"
  - Um (ou mais) números que indicam a "saúde" da empresa

- **Scorecards**
  - Metodologias para medir "o que é importante" num dado negócio
  - Técnicas para elaboração de "***balanced scorecards***"
- Identificar os **KPI** – Key Performance Indicator

21

## Acesso à datawarehouse

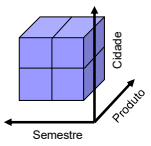
- Arquitectura em 3 níveis




22

## Sistemas de OLAP

- OLAP- On-Line Analytical Processing
  - Disponível para muitos sistemas de bases de dados
  - Conjunto de ferramentas de "reporting": fáceis e flexíveis
- Conceito de **hipercubo de dados**
  - Agrupar segundo **diversas dimensões**
    - Tempo, Local, Produto, Cliente, etc.
  - **Cortes** (slices) e **vistas**
    - Ver o hipercubo sob uma dada perspectiva
    - "Colapsar" (ou não) algumas dimensões
  - **Roll-up**:
    - Consolidar ou agregar em dados mais gerais
  - **Drill-down**:
    - Separar em nódulos mais específicos
  - Outras:
    - Ranking, Filtering, Dicing, estruturas ROLAP, HOLAP




23

## Exemplo de um cubo de dados

- dados de vendas por semestre, por produto e por cidade:

| *Semestre* | *Vendas* |
|---|---|
| Primeiro | 16.000,00 |
| Segundo | 16.000,00 |

| *Produto* | *Vendas* |
|---|---|
| Banana | 16.000,00 |
| Laranja | 16.000,00 |

| *Cidade* | *Vendas* |
|---|---|
| Lisboa | 16.000,00 |
| Porto | 16.000,00 |

24

## Exemplo de um cubo de dados

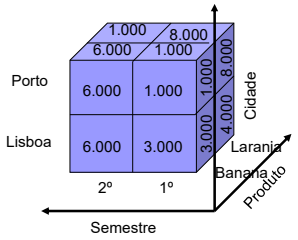
- Dados mais detalhados: numa tabela

| *Semestre* | *Produto* | *Cidade* | *Valor* |
|---|---|---|---|
| Primeiro | Banana | Lisboa | 3.000,00 |
| Primeiro | Banana | Porto | 1.000,00 |
| Primeiro | Laranja | Lisboa | 4.000,00 |
| Primeiro | Laranja | Porto | 8.000,00 |
| Segundo | Banana | Lisboa | 6.000,00 |
| Segundo | Banana | Porto | 6.000,00 |
| Segundo | Laranja | Lisboa | 3.000,00 |
| Segundo | Laranja | Porto | 1.000,00 |

25

## Exemplo de um cubo de dados

- Dados mais detalhados: num cubo




26

## Bibliografia

- George Marakas, Modern Data Warehousing, Mining, and Visualization, Prentice-Hall 2003
- Barry Devlin, Data Warehouse – from Architecture to Implementation, Addison-Wesley, 1997

27

## Bibliografia (artigos)

- Kouzes et al., The changing paradigm of Data-Intensive computing, IEEE Computer, Jan 2009

28