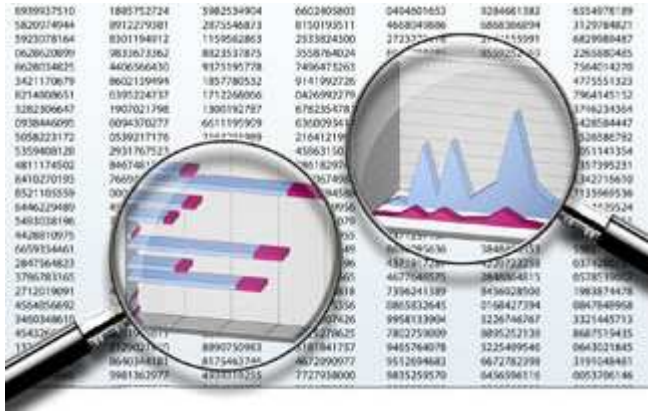


ITnerante

TIMASTERS

Banco de dados

Suporte a decisão – Business Intelligence

Curso Preparatório - ITnerante

Prof. Thiago Cavalcanti

Ementa

- Módulo 02 – Soluções de suporte à decisão
  - Recuperação e visualização de dados
    - OLAP
    - Data Mining
    - Painéis e Dashboards
  - Qualidade de dados

ITnerante

Tire suas dúvidas

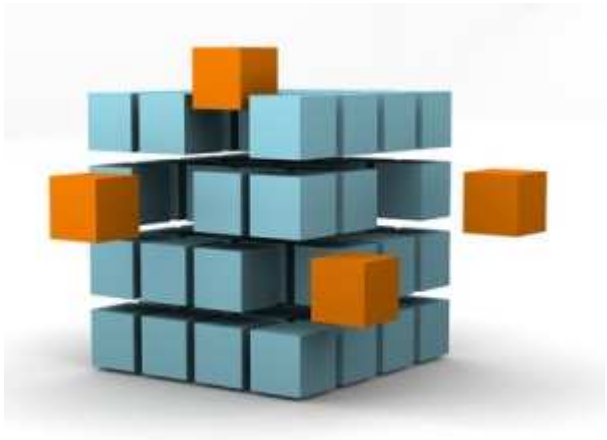
- rcthiago@gmail.com
- www.itnerante.com.br
- Lista: timasters@yahoogroups.com.br

ITnerante

TIMASTERS

Recuperação e visualização de dados

Prof. Thiago Cavalcanti



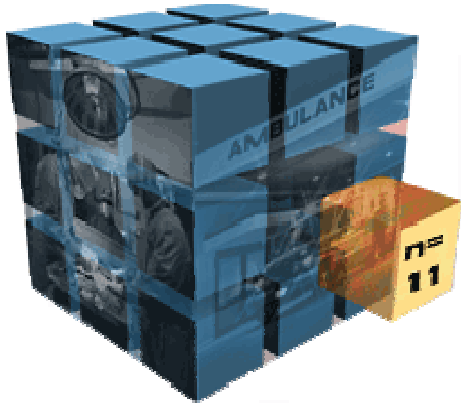
ITnerante

TIMASTERS

OLAP

On-Line Analitical Processing

(Processamento Analítico On-Line)



Prof. Thiago Cavalcanti

Sopa de letrinhas

O que é OLAP?

- **Processamento** de dados dedicado ao suporte a decisão
  - Por meio de visualização de dados agregados ao longo de várias dimensões analíticas (ex.: tempo, espaço, categoria de produto, quantidade vendida, preço...)
  - Armazenados em BD especializadas
    - Seguem um modelo lógico de dados multidimensional
    - Chamados de Data Warehouse, Data Mart ou BD multidimensionais
  - Hierarquizadas em várias granularidades

Conceitos de OLAP

- Laudon&Laundon: Recurso que permite **manipular e analisar** grandes volumes de dados sob **múltiplas perspectivas**.

- Webopedia: Uma categoria de **ferramentas de software** que prover **analise de dados** armazenados em um banco de dados.
  - Permite analise de **diferentes dimensões** dos dados **multidimensionais**

Organizando as coisas

OLAP x OLTP

Operações OLAP

Taxonomia

Comparativo OLTP x OLAP

Características	OLTP	OLAP
Operação típica	Atualização	Análise
Telas	Imutáveis	Definida pelo Usuário
Nível de dados	Atomizado	Altamente Sumarizado
Recuperação	Poucos Registros	Muitos registros
Orientação	Registros	Arrays
Modelagem	Processo / Aplicação	Assunto
Natureza dos dados	Permite atualizações contínuas	Dados históricos, sumariados e integrados

Comparativo OLTP x OLAP (Turban)

Características	OLTP	OLAP
Propósito	Dar suporte ao dia-a-dia operacional da empresa	Dar suporte a tomada de decisão e prover respostas para as consultas de negócio e gerenciamento
Fonte de dados (Data Source)	Transacional	Data warehouse ou data mart
Relatórios	Rotineiros, periódicos, relatórios pontuais (focados)	Ad hoc, Multidimensionais, relatórios e consultas de larga amplitude
Tempo de execução	Possibilita processamento mais **eficiente** de transações (rápido)	Possibilitar processamento mais eficiente para apresentação de dados focados na tomada de decisão (lento)

Comparativo OLTP x OLAP (Barbieri)

CARACTERÍSTICAS	DADOS OPERACIONAIS	DADOS INFORMACIONAIS
1. Conteúdo	Valores correntes	Valores sumarizados, calculados, integrados de várias fontes
2. Organização dos dados	Por aplicação/sistema de informação	Por assuntos/negócios
3. Natureza dos dados	Dinâmica	Estática até o *refreshment* dos dados, de tempos em tempos
4. Formato das estruturas	Relacional, próprio para computação transacional	Dimensional, simplificado, próprio para atividades analíticas
5. Atualização dos dados	Atualização campo a campo	Acesso granular ou agregado, normalmente sem *update* direto
6. Uso	Altamente estruturado em tabelas, processamento repetitivo	Estruturado em fatos e dimensões, com processamento analítico/preditivo
7. Tempo de resposta	Otimizado para faixas abaixo de 1 seg	Análises mais complexas, com tempos de respostas maiores

Questão 01. CESGRANRIO - 2012 - PETROBRAS - Analista de Sistemas Júnior - Engenharia de Software

Uma característica típica de uma aplicação OLAP é:

(a) focar as consultas sobre dados brutos.

(b) manipular principalmente dados atuais.

(c) recuperar pequenas quantidades de dados por consulta.

(d) ser orientada para arrays de dados.

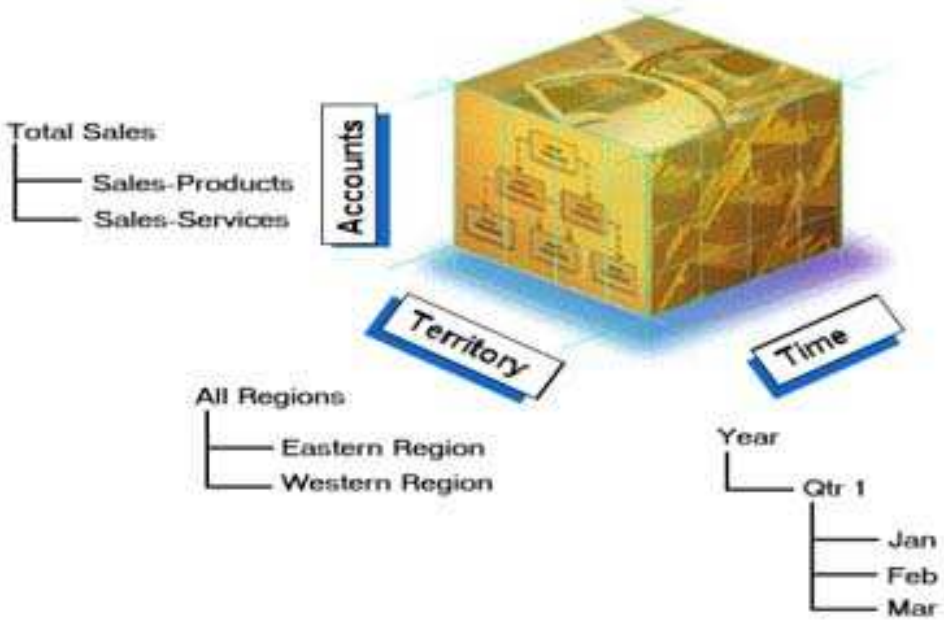
(e) utilizar basicamente consultas predefinidas.

Questão 02. CESPE - 2013 - MPU - Analista - Desenvolvimento de Sistemas

Julgue os itens a seguir, a respeito de soluções de suporte à decisão.

[119] Uma característica distinta dos data warehouses é o seu direcionamento para aplicações de apoio às decisões. Eles são otimizados para a recuperação de dados, não para o processamento rotineiro de transações.

[120] OLAP (online analytical processing) é um termo utilizado para descrever a análise de dados complexos a partir do data warehouse. As ferramentas OLAP empregam as capacidades de computação distribuída para análises que requerem mais armazenamento e poder de processamento que as disponibilizadas por um desktop.



Funções ou funcionalidade de OLAP

SLICE AND DICE, PIVOT, DRILL DOWN, DRILL UP, ROLL UP, DRILL THROUGHT, DRILL ACROSS

Granularidade

- A granularidade de dados refere-se ao **nível de sumarização** dos elementos e de detalhe disponíveis nos dados
  - Considerado, por alguns estudiosos, **o mais importante aspecto** do projeto de um Data Warehouse.

OLAP engine

- OLAP oferece recursos de modelagem analítica, incluindo um mecanismo de cálculo para desvio padrão, variância, etc. , e **processamento de medidas em múltiplas dimensões**
- **Gera sumarizações**, **agregações** e **hierarquias** em cada **nível de granularidade** e em cada cruzamento de dimensão
- Suporta modelos funcionais para previsão, análise de tendências e análise estatística
- Neste contexto, **um motor OLAP** é
  - **Uma ferramenta poderosa a análise de dados**

Questão 03. TRT - 1ª REGIÃO (RJ) - Analista Judiciário Tecnologia da Informação - 2011

59. Ao nível de sumarização dos elementos e de detalhes disponíveis nos dados em um DW dá-se o nome de

(a) relacionamento.

(b) capacidade.

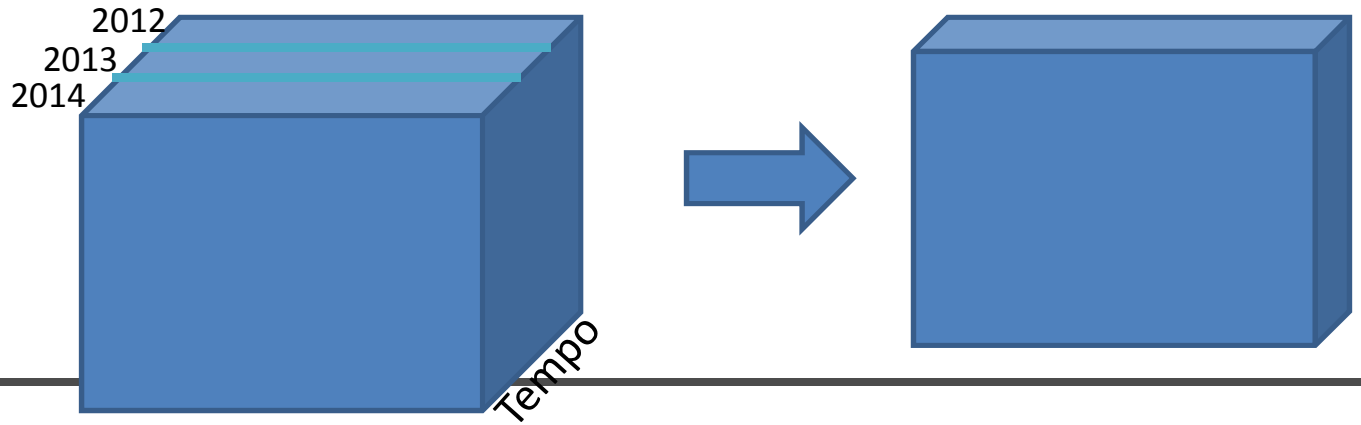
(c) granularidade.

(d) integridade.

(e) arquitetura.

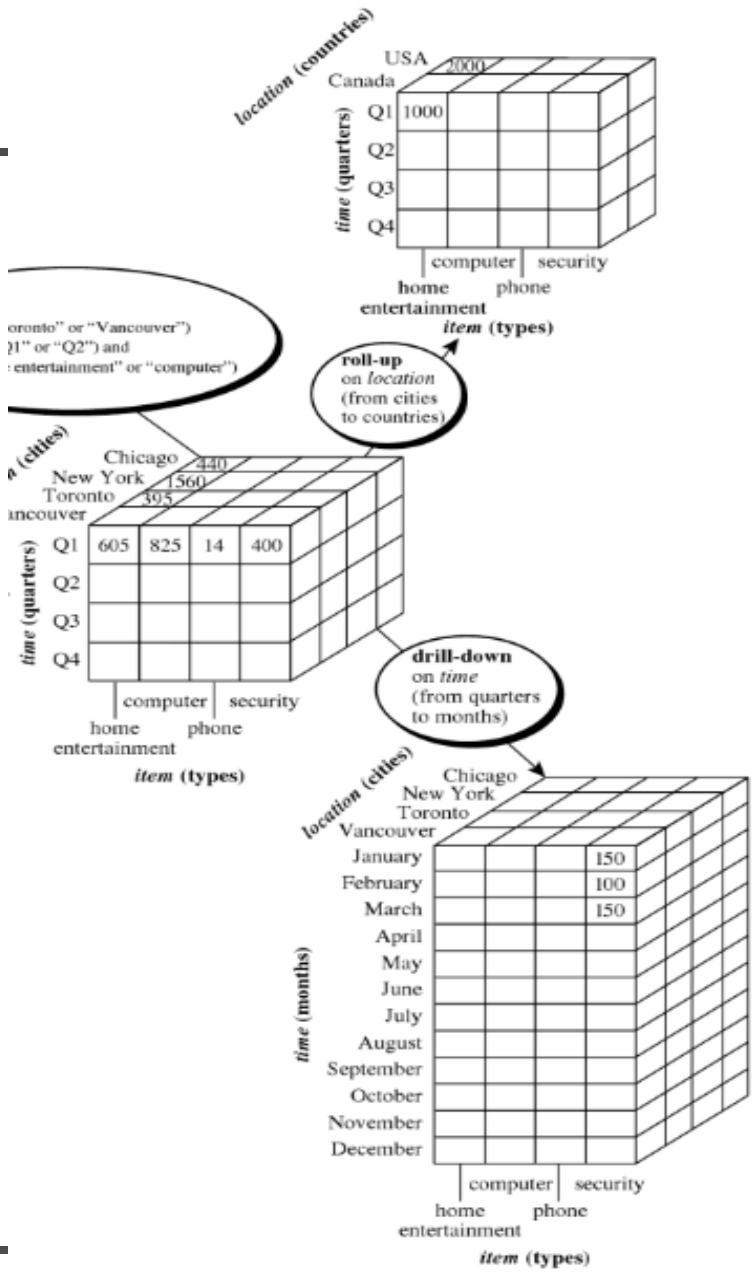
Drill Up ou Roll up

- Aplica uma agregação sobre o cubo de dados
- Aumenta o nível de granularidade
  - **Subindo** na hierarquia da dimensão
  - Realiza **uma redução** sobre a dimensão
- **Diminuindo o nível de detalhamento** da informação



Drill down

- É o inverso do roll-up
- **Aumenta** o **nível de detalhe** da informação
  - Navega de um dado menos detalhado para um mais detalhado
  - Realizado por uma decida na hierarquia da dimensão ou
  - Por meio da introdução de uma nova dimensão
- Diminui nível de granularidade.



Roll-up

Drill-down

ITnerante

Questão 04. TRT - 23ª REGIÃO (MT) - Analista Judiciário - Tecnologia da Informação - 2011

Q.56.A funcionalidade pré-programada de resumir os dados, com generalização crescente, oferecida pelas aplicações por meio das ferramentas de construção de *data warehouses* é denominada

(a) *Roll up.*

(b) *Drill down.*

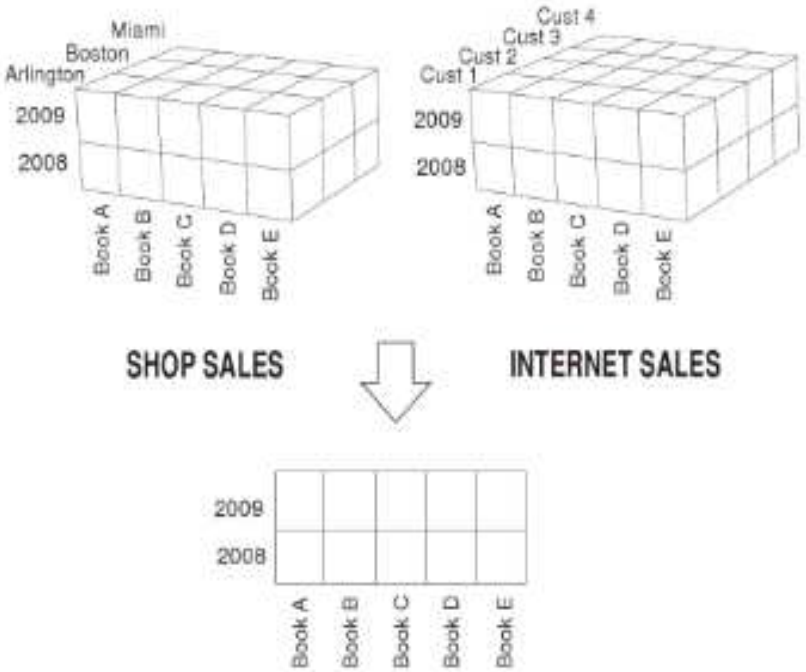
(c) *Pivot.*

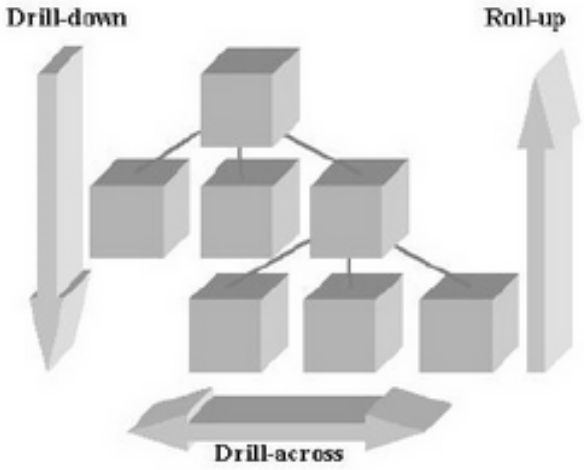
(d) *Sorting.*

(e) *Slice and dice.*

Drill Across

- Executa uma consulta envolvendo **mais de uma tabela fato**!
- Essa operação exige que **os dois cubos** tenham pelo menos **uma das dimensões em comum**.
- A ideia é você conseguir consultar as múltiplas tabelas fato e colocar o resultado em um único *data set*.

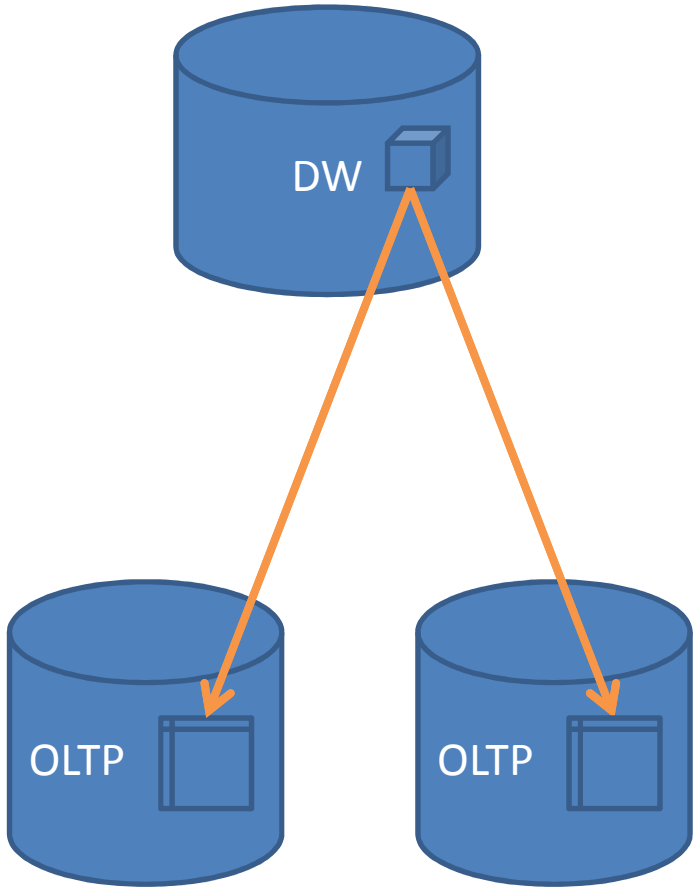




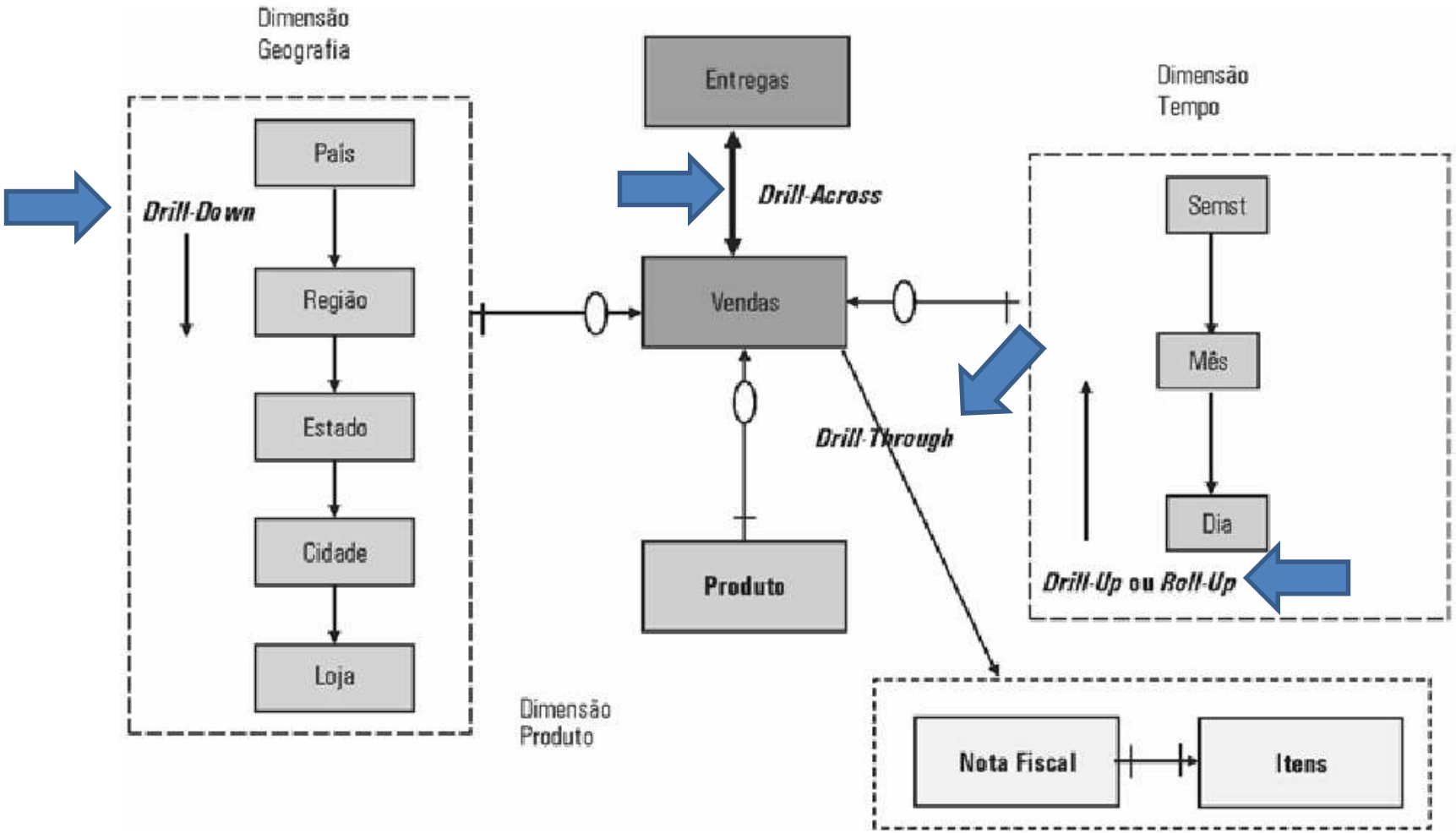
ITnerante

Drill Throught

- As tabelas de *drill-through* permitem que você exiba, em tempo de consulta, **os detalhes dos dados não sumarizados**
  - A partir dos quais uma célula de uma tabela ou uma seleção de células é sumarizada
- Permite as empresas acesso aos dados que não estão armazenados no servidor OLAP, fazendo-as acessíveis para os usuários finais das aplicações OLAP
  - Esses dados podem vir tanto do **DW** quanto das **bases transacionais.**



Resumindo os DRILLs

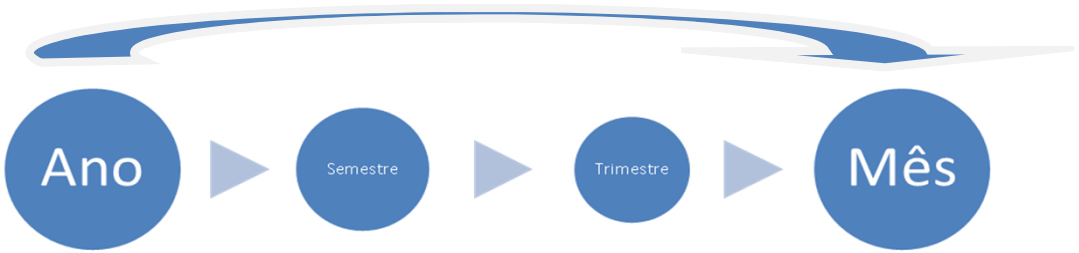


ITnerante

Drills Across e Drill throught (Definição 02)

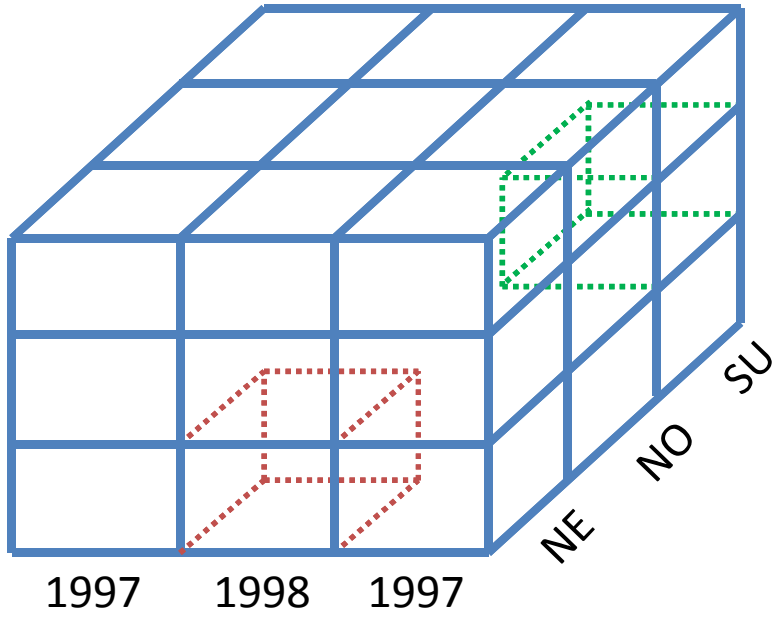
- **Drill Across -** ocorre quando o usuário **pula um nível intermediário** dentro de uma mesma dimensão.
  - Por exemplo, a dimensão tempo é composta por ano, semestre, trimestre, mês e dia.
  - A operação Drill Across é executada quando o usuário passa **de ano direto para trimestre ou mês**.
- **Drill Throught** - ocorre quando o usuário passa de uma informação contida em uma dimensão para uma outra.
  - Por exemplo: Inicia na dimensão do tempo e no próximo passo analisa a informação por região

Drill Across e Drill Throught



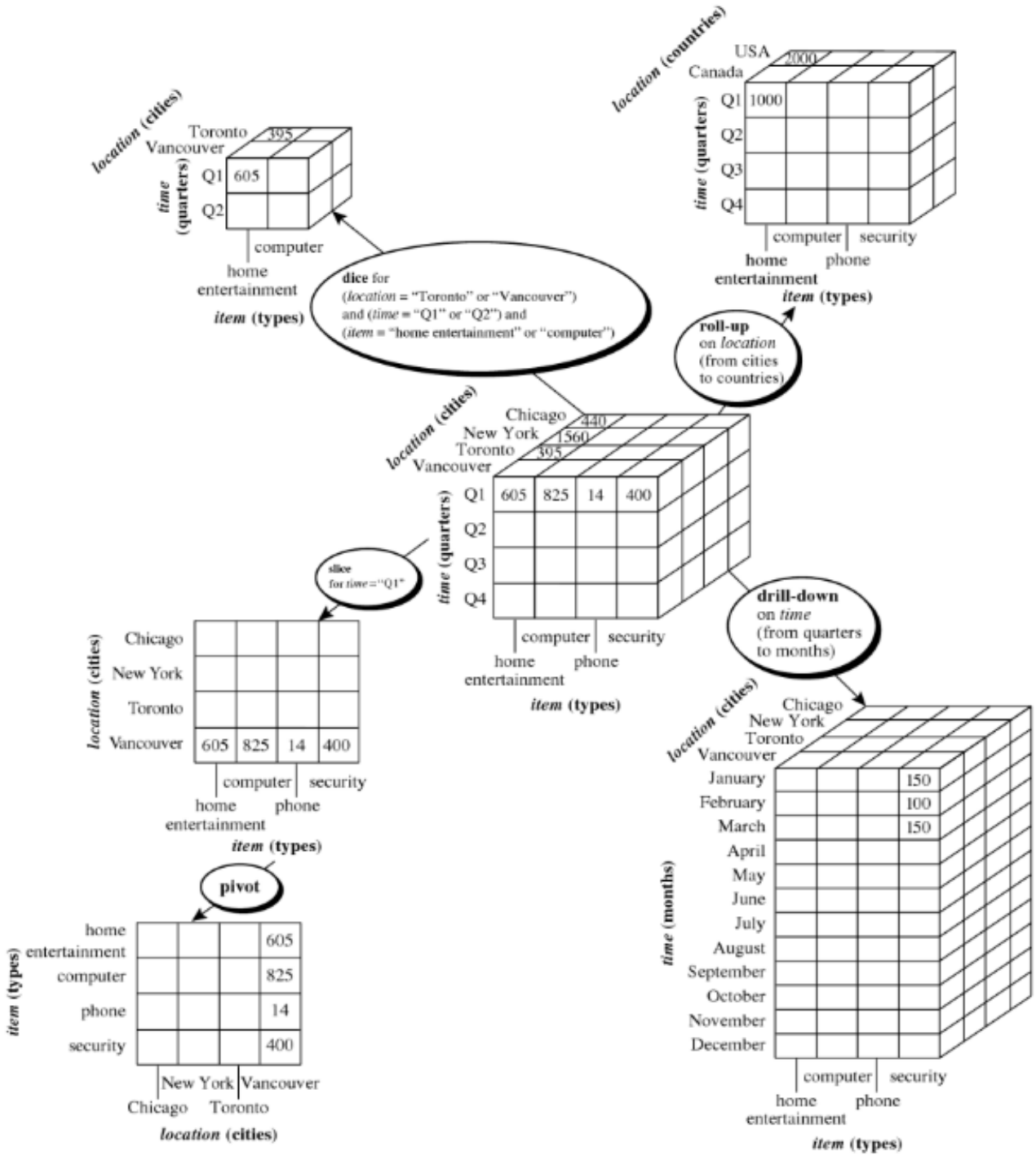
Drill Across

Drill Throught



ITnerante

Slice & Dice
Pivot

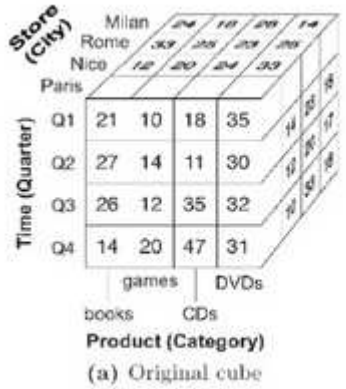


ITnerante

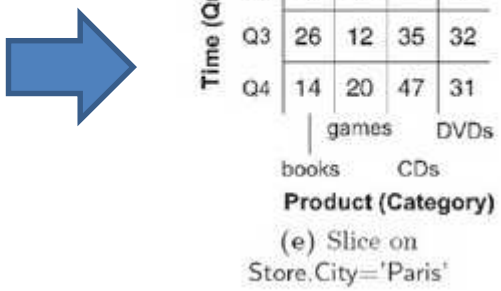
Slice and Dice – Turban (Definição 01)

- **Slice** - é um subconjunto de uma matriz multidimensional (geralmente bidimensional) correspondendo a um único valor definido para um (ou mais) das dimensões no subconjunto.
- **Dice** – é um slices em mais de uma dimensão de um cubo de dados

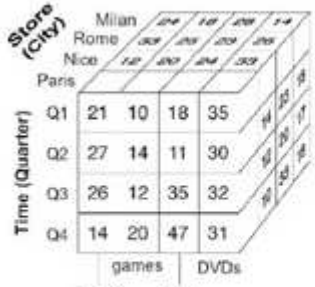
Slice and Dice (Definição 01)



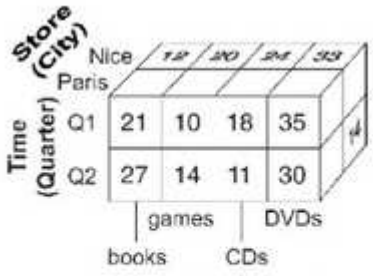
(a) Original cube



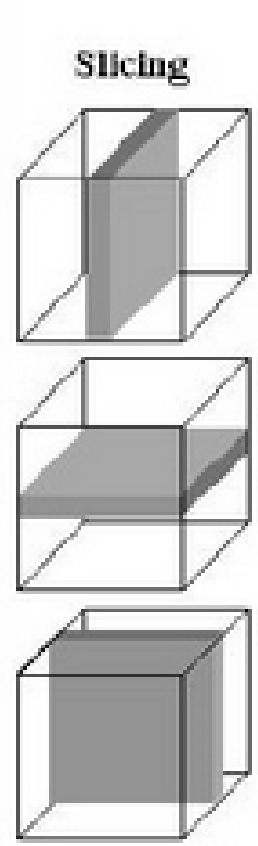
(e) Slice on Store.City='Paris'



(a) Original cube



(f) Dice on Store.Country='France' and Time.Quarter='Q1' or 'Q2'



ITnerante

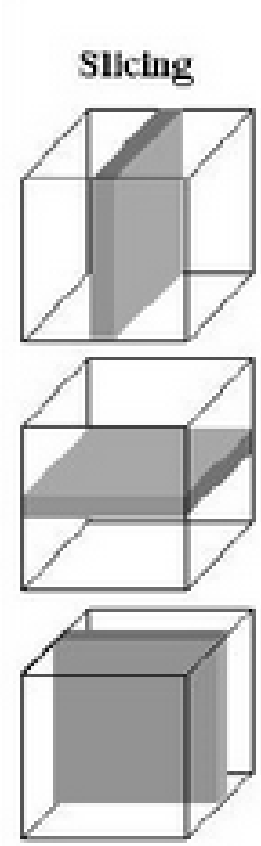
Slice and Dice (Definição 02)

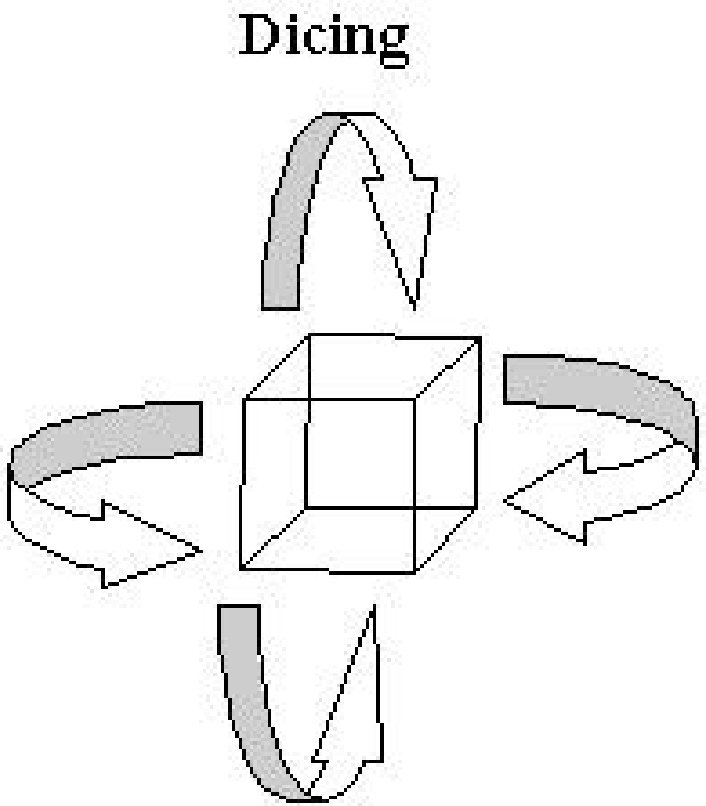
- **Slice and Dice** - é uma das principais **características** de uma ferramenta OLAP.
- Como a ferramenta OLAP recupera o microcubo?
  - No OLAP, as informações são armazenadas em **cubos multidimensionais**, que gravam valores **quantitativos e medidas**, permitindo visualização através de **diversas perspectivas**.
  - Estas medidas são organizadas em **categorias descritivas**, chamadas de **dimensões** e formam, assim, a estrutura do cubo

Slice and Dice (Definição 02)

- Surgiu a necessidade de criar um módulo, que se convencionou de **Slice and Dice**, para ficar responsável por trabalhar a informação
  - Ele serve para **modificar a posição** de uma informação, **trocar linhas por colunas** de maneira a **facilitar a compreensão** dos usuários e **girar o cubo** sempre que tiver necessidade.

Slice and Dice (Definição 02)





Slice and dice (Definição 02)

- Estes tipos de navegação iniciada pelo usuário através dos dados são feitos por meio da especificação de slices (via rotações) e drill down/up (através de agregação) são às vezes chamados de "**slice and dice**"
- Operações OLAP comumente usados incluem slice and dice, drill-down, roll-up, e pivô.

Questão 05. FGV - 2010 - BADESC - Analista de Sistemas - Desenvolvimento de Sistemas

[36] OLTP - Online Transaction Processing é uma ferramenta de banco de dados e de Business Intelligence, utilizada para apoiar as empresas na análise de suas informações, com o objetivo final de transformar dados em informações capazes de dar suporte às decisões gerenciais de forma amigável e flexível ao usuário e em tempo hábil. No OLAP – Online Analytical Processing, as informações são armazenadas em cubos multidimensionais, que gravam valores quantitativos e medidas, permitindo visualização por meio de diversos ângulos. Estas medidas são organizadas em categorias descritivas, chamadas de dimensões e formam a estrutura do cubo. A respeito do OLAP, analise as afirmativas a seguir.

I. Drill Across: ocorre quando o usuário pula um nível intermediário dentro de uma mesma dimensão.

II. Slice and Dice: é uma das principais características de uma ferramenta OLAP e serve para modificar a posição de uma informação, trocar linhas por colunas de maneira a facilitar a compreensão dos usuários e girar o cubo sempre que tiver necessidade.

III. Drill Up: ocorre quando o usuário aumenta o nível de detalhe da informação, diminuindo a granularidade, ou seja, quais os tipos de consultas que podem ser feitas no DW, influenciando diretamente na velocidade do acesso às informações e no volume de dados armazenados.

Assinale:

(A) se somente a afirmativa I estiver correta.

(B) se somente as afirmativas I e II estiverem corretas.

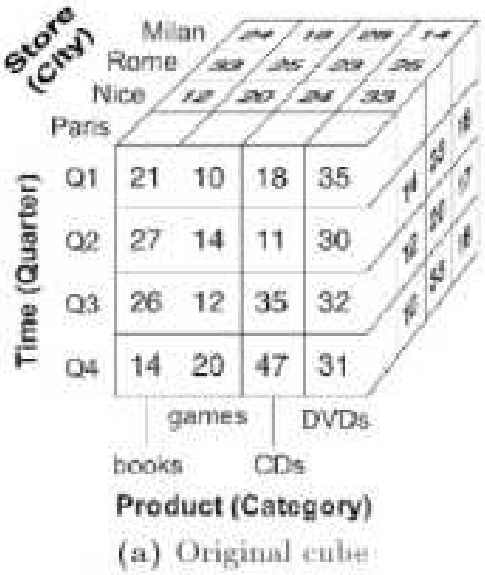
(C) se somente as afirmativas I e III estiverem corretas.

(D) se somente as afirmativas II e III estiverem corretas.

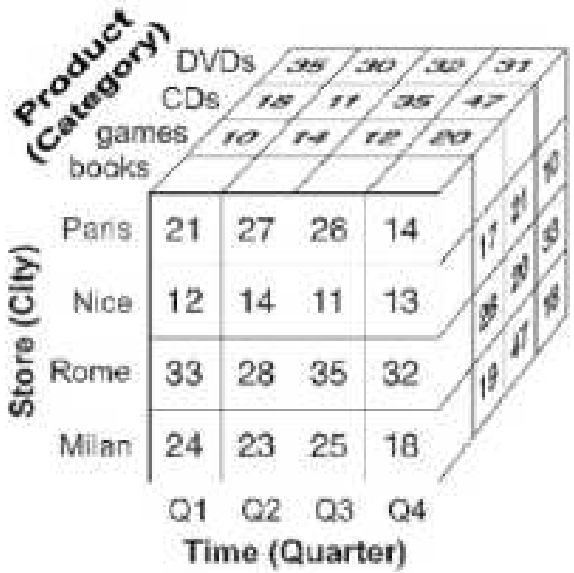
(E) se todas as afirmativas estiverem corretas.

Pivot ou Pivotiamento

- A operação pivot (ou rotate) roda os eixos de um cubo para oferecer uma alternativa de visualização dos dados



(a) Original cube



(d) Pivot

ITnerante

Questão 06. FCC - 2014 - TRF - 3ª REGIÃO - Analista Judiciário - Informática - Banco de Dados

43. A tecnologia de *Data Warehouse* oferece suporte às ferramentas OLAP, que apresentam visões multidimensionais de dados permitindo a análise das operações de negócio para facilitar a tomada de decisões. Estas ferramentas suportam algumas operações de maneira a dar aos analistas o poder de observar os dados de várias maneiras em níveis diferentes. Considere duas destas operações mostradas nas figuras abaixo.

Ano	Dados	Região			
		Ásia	Europa	América do Norte	Total Geral
2010	Soma de *Hardware*	97	23	198	318
	Soma de *Software*	83	41	425	549
2011	Soma de *Hardware*	115	28	224	367
	Soma de *Software*	78	65	410	553
2012	Soma de *Hardware*	102	25	259	386
	Soma de *Software*	55	73	497	625
Soma de *Hardware* Total		314	76	681	1071
Soma de *Software* Total		216	179	1332	1727

Região	Dados	Ano			
		2010	2011	2012	Total Geral
Ásia	Soma de *Hardware*	97	115	102	314
	Soma de *Software*	83	78	55	216
Europa	Soma de *Hardware*	23	28	25	76
	Soma de *Software*	41	65	73	179
América do Norte	Soma de *Hardware*	198	224	259	681
	Soma de *Software*	425	410	497	1332
Soma de *Hardware* Total		318	367	386	1071
Soma de *Software* Total		549	553	625	1727

Figura 1

Região	Variação de vendas
Africa	105%
Ásia	57%
Europa	122%
América do Norte	97%
Pacífico	85%
América do Sul	163%

País	Variação de vendas
China	123%
Japão	52%
Índia	87%
Cingapura	95%

Figura 2

As operações mostradas na Figura 1 e na Figura 2, respectivamente, são

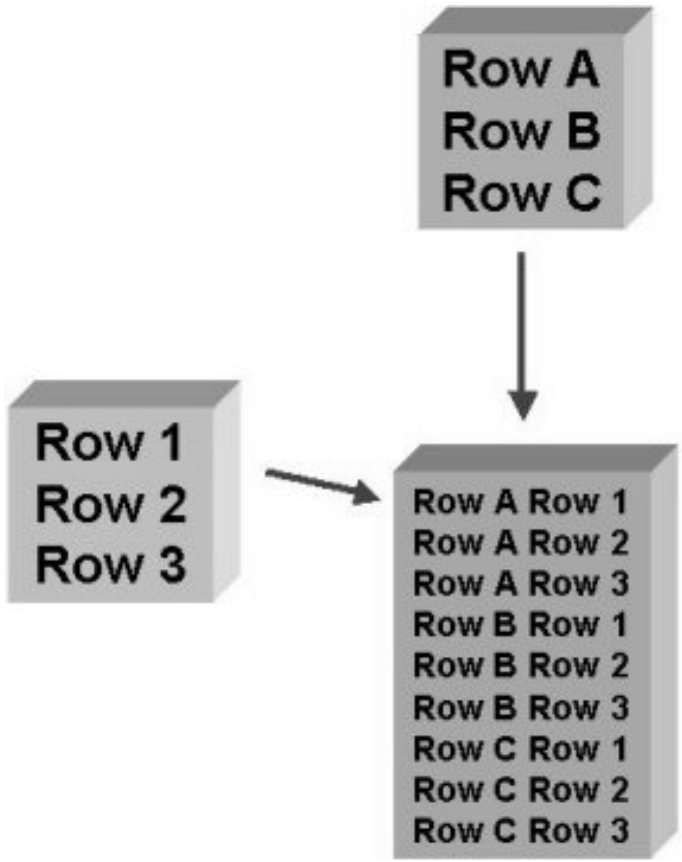
(A) *drill-down* e ROLAP.
(B) rotação e *drill-down*.
(C) ROLAP e *drill-through*.
(D) rotação e *roll-up*.
(E) *roll-up* e rotação.

Outros comandos

- Algumas ferramentas possuem um conjunto muito variado de operadores dimensionais, estatísticos e temporais. As mais comuns são:
  - RANKING: Classifica determinada informação baseada nos *n* melhores indicadores (top-N, bottom-N)
  - LAST-WEEK: Mostra os valores relacionados à semana anterior, tendo como referência a semana atual
  - PRIOR-WEEK: Somente os valores relacionados ao período compreendido nos últimos sete dias, tendo como referência a data atual.
  - YEAR-TO-DATE: Compreendendo o período do ano de referência até a data de hoje.

Produto cartesiano (Cross-join)

- É usado para gerar um produto cartesiano entre os conjuntos passados como parâmetro.



MDX - Cross-join (*)

```
SELECT
 {
        [Measures].[Sales Amount - Fact Reseller Sales]
 }
 ON COLUMNS,
 NON EMPTY {[Dim Reseller].[Business Type].members *
 [Dim Geography].[English Country Region Name].members}
 DIMENSION PROPERTIES MEMBER_NAME ON ROWS
 FROM [Adventure Works DW]
CELL PROPERTIES VALUE, FORMATTED_VALUE, FORE_COLOR, BACK_COLOR
```

Questão 07. TRT - 11ª Região (AM) - Analista Judiciário - Tecnologia da Informação – 2010

37. No âmbito dos DWs e OLAP, o processo onde se faz a junção dos dados e transforma-se as colunas em linhas e as linhas em colunas, gerando dados cruzados, é chamado de

(a) drill-across.

(b) star.

(c) cube.

(d) pivot.

(e) cross-join.

Reflexão

- Quando um usuário realiza um drill-down, ele está, no fundo, solicitando um SQL.
- Esse SQL irá modificar o seu GROUP BY, obtendo o dado na granularidade solicitada.

Facilidades de SQL

- Algumas facilidades para dar suporte a OLAP foram incluídas no SQL 1999 e outras foram inseridas posteriormente.
  - Essencialmente o GROUPING SETS, ROLLUP e o CUBE que são extensões do GROUP BY.
  - Novas funções numéricas(natural logarithm, exponentiate, power, square root ...)
  - Novos operadores de agregação (variance, standard deviation)
  - Ranking functions
  - ...

Exemplo

```
SELECT p.produto#, d.trimestre, SUM(v.faturamento)
FROM vendas v, data d, Produto p
WHERE v.produto# = p.produto#
AND v.data = d.data#
AND d.trimestre IN (‘2’, ‘3’)
AND d.ano = 1999
AND p.produto# IN (‘P1’, ‘P2’)
GROUP BY {CUBE,ROLLUP} (p.produto#, t.trimestre)
```

Questão 08. TRE-PE - Analista Judiciário - Análise de Sistemas - 2011

Q.39.No âmbito das agregações SQL em data warehouse (DW), cube e rollup são extensões da cláusula

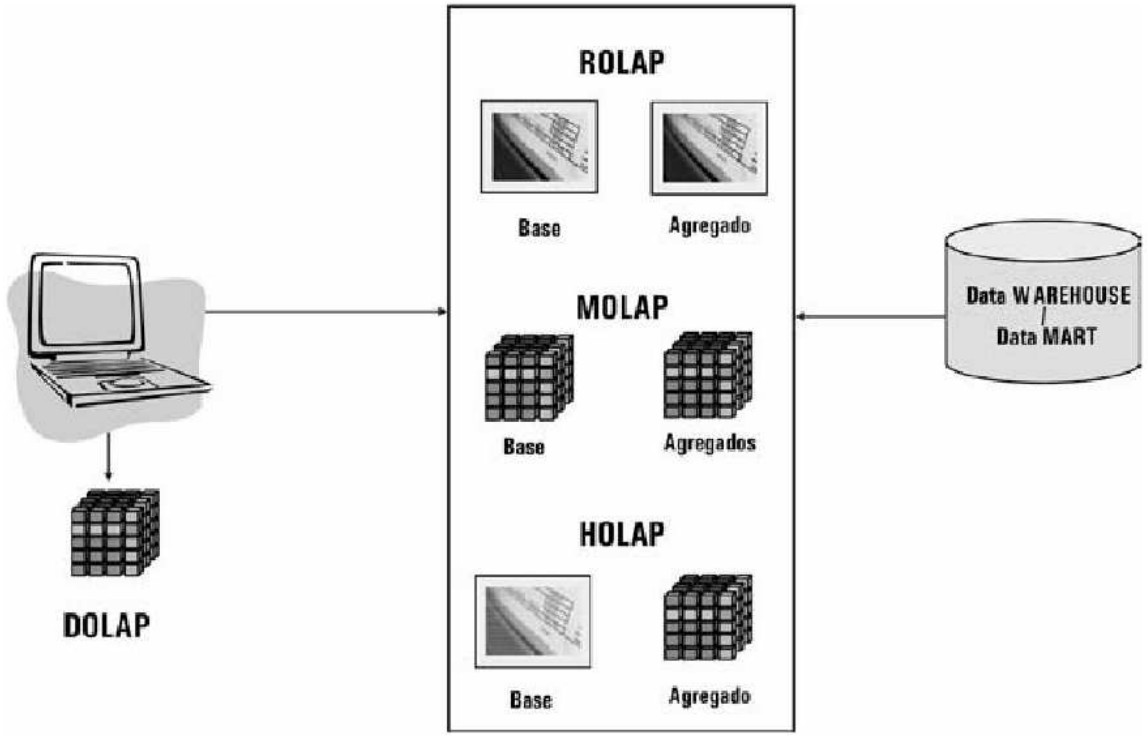
(a) *having.*

(b) *group by.*

(c) *avg.*

(d) *grouping sets.*

(e) *order by*



ROLAP/MOLAP/HOLAP/DOLAP ...

ARQUITETURA DE SERVIDORES OLAP

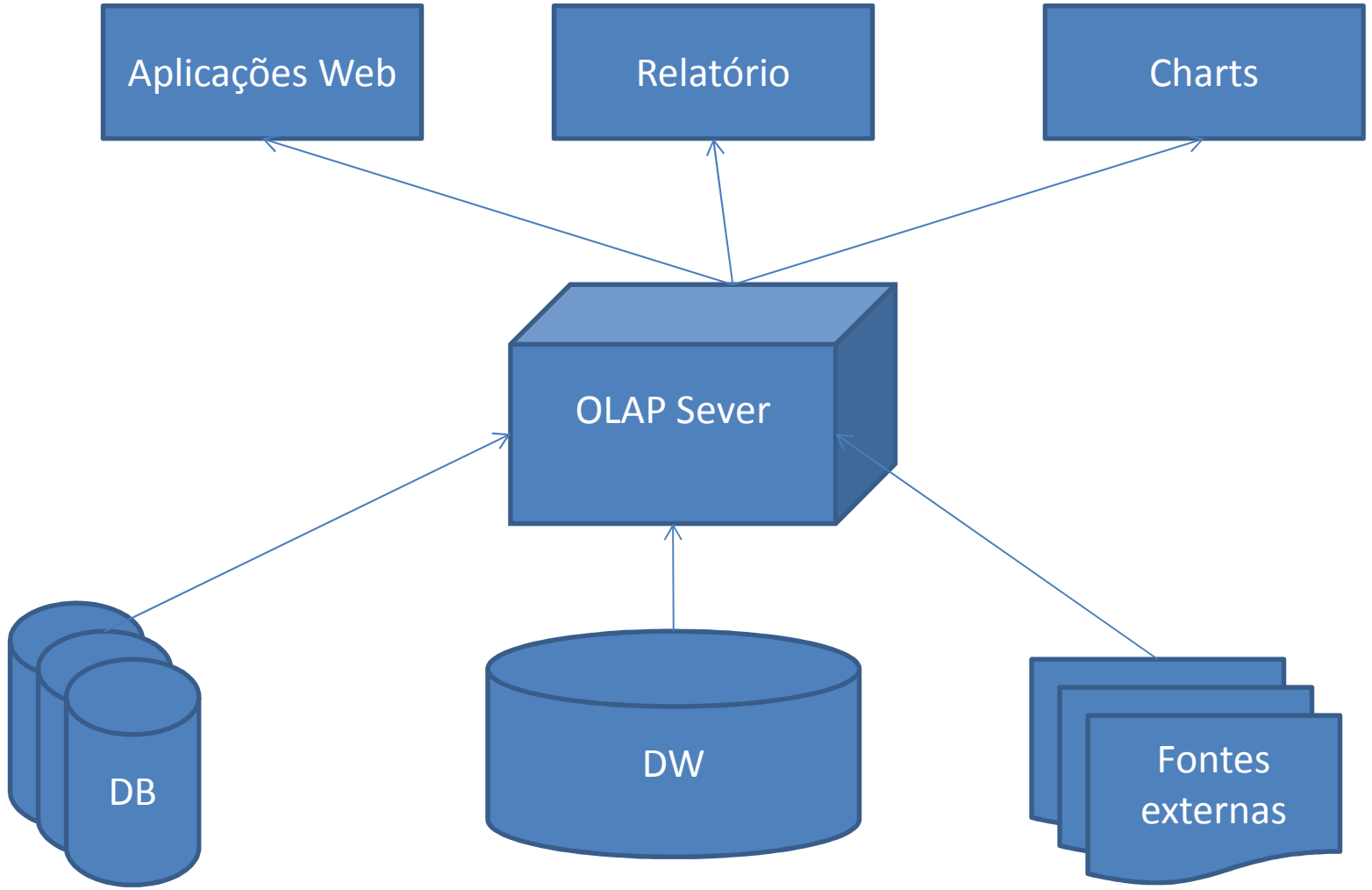
Arquitetura de Servidores OLAP

- Logicamente, servidores OLAP apresentam aos usuários de negócio os dados multidimensionais de um Data Mart ou de um Data Warehouse, sem a preocupação de mostrar como e onde os dados são de fato armazenados
- Todavia, a arquitetura física e a implementação de servidores OLAP **devem considerar aspectos de armazenamento**
- A implementação de um DW para processamento OLAP pode ser feita das seguintes formas:
  - ROLAP, MOLAP, HOLAP, ...

Arquitetura OLAP

- Componentes
  - Data source - **as fontes de dados** usadas para a analise OLAP.
  - OLAP Server - link entre o banco dados e o cliente. Gerencia a estrutura de dados multidimensional.
  - OLAP Customer - são aqueles que fornecem aplicações de mineração de dados, mas também suportam a geração de resultados (graphs, reports, etc.).

Arquitetura



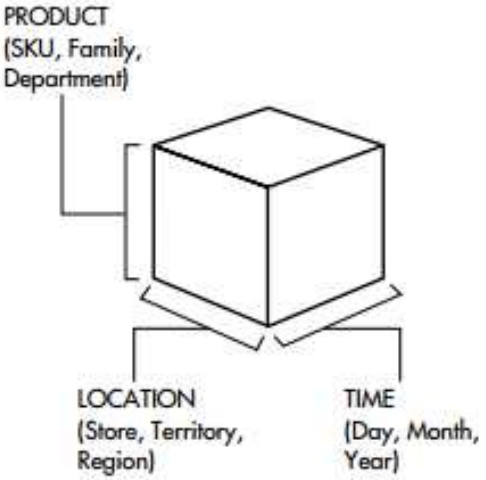
Servidores OLAP (MDDB x RDB)

- O **back-end** de um sistema OLAP é **o servidor OLAP**
- Faz todo o processamento (dependendo do modelo do sistema), é nele que os dados efetivamente acessados são armazenados.
- Diferentes filosofias governam a arquitetura do servidor, uma das principais características de um produto OLAP é se o servidor usa
  - **Um banco de dados multi-dimensional** (MDDB) para armazenar os dados, ou
  - Um **banco de dados relacional** (RDB).

Dados Agregados/Pré-agregados

"Quanto maior a necessidades de cálculos para produzir um conjunto de informação, maior será o tempo de resposta!"

- Pedaços de informação que são frequentemente acessados devem ser pré-agregados.
  - São portanto, pré-calculados e armazenados como um novo dado dentro da base.
  - Por exemplo: Vendas por mês, ...



Multidimensional Databases
Multidimensional Databases have singular, hierarchical data access paths, effectively limiting the number of attributes available for user queries.

MOLAP

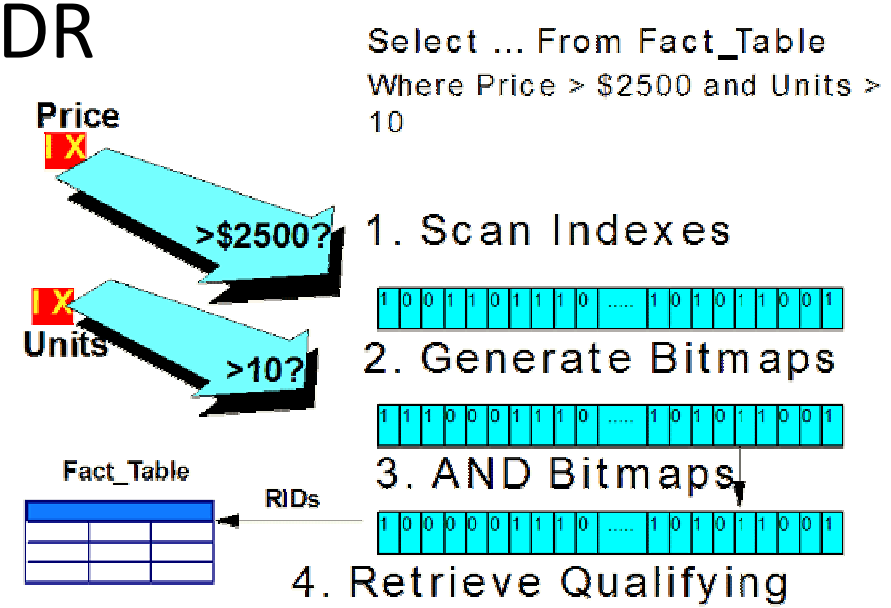
MOLAP

- Multidimensional On-Line Analytical Processing.
  - Isto significa que o servidor usa um MDDB para armazenar dados.
  - Estratégia pela qual são usados gerenciadores de banco de dados proprietários, com características de armazenamento especiais e ferramentas para tratamento dimensional de dados.

MOLAP – Multidimensional OLAP Server

- Dispõe de propriedades especiais de armazenamento como **matrizes esparsas**, **operações com array** e **indexações de bitmap**
- Não oferece toda a gama de recursos encontradas num SGBDR

$$A = \begin{pmatrix} 50 & 0 & 0 & 0 \\ 10 & 0 & 20 & 0 \\ 0 & 0 & 0 & 0 \\ -30 & 0 & -60 & 5 \end{pmatrix}$$



ITnerante

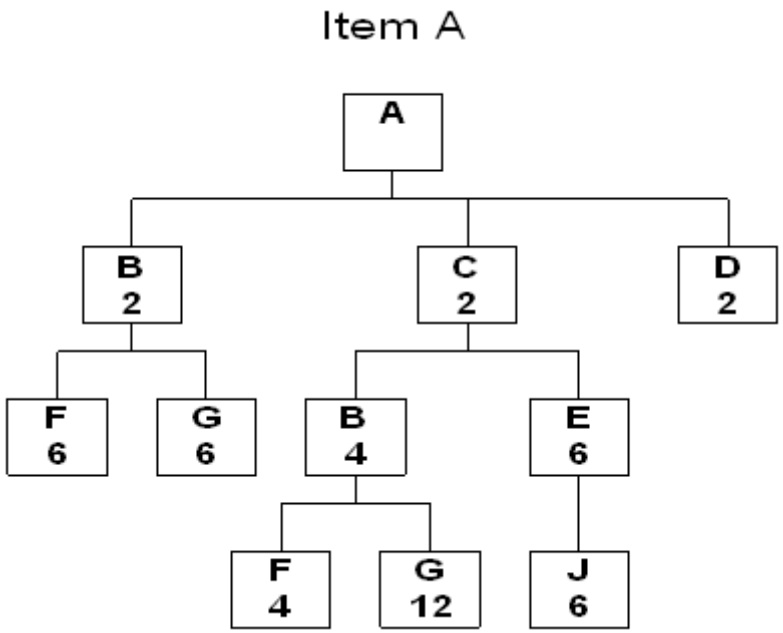
Bancos de dados Multidimensionais

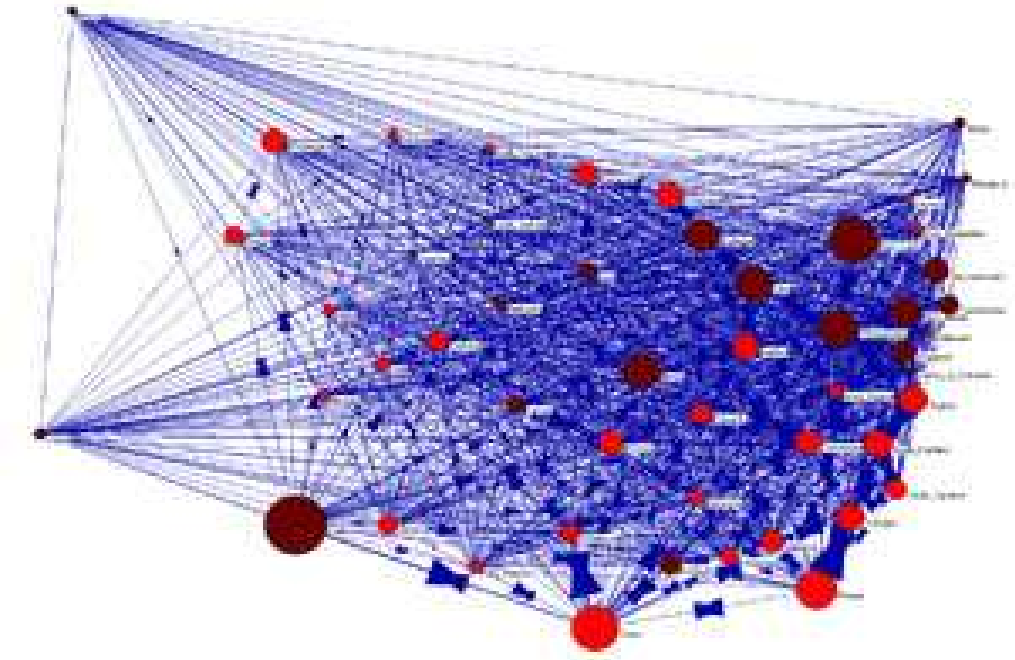
Pros	Contras
Preciso para modelo de dados de negócio	Não consegue gerenciar grandes bancos de dados (VLDB)
Acesso rápido **sem SQL**	Nova tecnologia **não totalmente otimizada**
Dados sumarizados **pré-calculados**	Risco de **explosão** de banco de dados

Database Explosion

- É um fenômeno em MDDB.
  - Geralmente está relacionado com a dispersão de base de dados e “pré-agregação” dos dados.
  - Se um banco de dados multidimensional contém um pequeno número de pontos de dados em comparação com o número de níveis de agregação, cada pedaço de dados terá uma maior contribuição para todos os dados obtidos a partir dele.
    - Quando a base de dados "explode", o tamanho da base de dados se torna de magnitude maior do que deveria ser.

Abstração





ITnerante

Database Explosion

- É difícil determinar condições para a explosão de dados, ou para prever se um configuração particular vai explodir.
- Uma abordagem que parece ajudar a resolver o problema é a manipulação de dados esparsos dinamicamente.
  - Manipulação de dados esparsos de forma dinâmica permite que um banco de dados analisar seus padrões de armazenamento próprios e otimizá-los para evitar a explosão de dados.

Questão 09. PGE-RJ - Técnico Superior de Análise de Sistemas e Métodos - 2009

MOLAP é

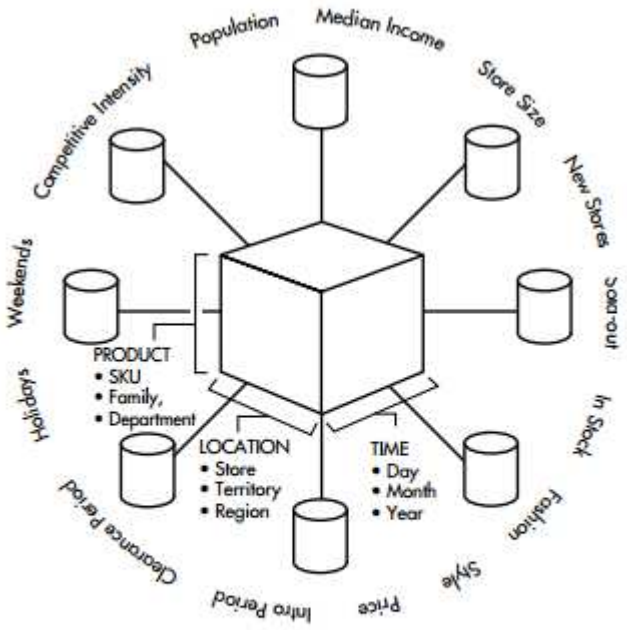
(a) utilizado para análise de segurança e usabilidade de dados em bancos relacionais.

(b) um instrumento utilizado no *tuning* de bancos de dados.

(c) uma ferramenta de monitoração de redes de computadores.

(d) uma ferramenta de proteção de redes de computadores.

(e) um mecanismo utilizado no âmbito dos bancos de dados multidimensionais.



Relational Databases
Relational Databases, with their support for joins, permit almost limitless, user-oriented access paths to data.

ROLAP

ROLAP

- O termo ROLAP especifica que o servidor OLAP baseia-se numa base de dados relacional.
  - Relational On-Line Analytical Processing.
- Os dados de origem são inseridos em um banco de dados relacional, geralmente em **um esquema estrela ou esquema floco de neve**, o que ajuda em tempos de recuperação rápidos
- O servidor fornece um modelo multidimensional dos dados, através de **consultas SQL otimizadas**

Razões para escolher ROLAP

- RDBs são uma tecnologia bem estabelecida que tem tido muitas oportunidades para otimização.
- Suporta maior quantidade de dados que uma MDDB.
  - São construídos para isso!

ROLAP – Relacional OLAP Server

- Esse é **um servidor intermediário** que fica entra a base de dados relacional de **back-end** e as ferramentas de **front-end**
- Eles usam SGBDs relacionais ou relacionais estendidos para gravar e gerencias os dados do DW, e um middleware OLAP para dar suporte às funcionalidades faltantes

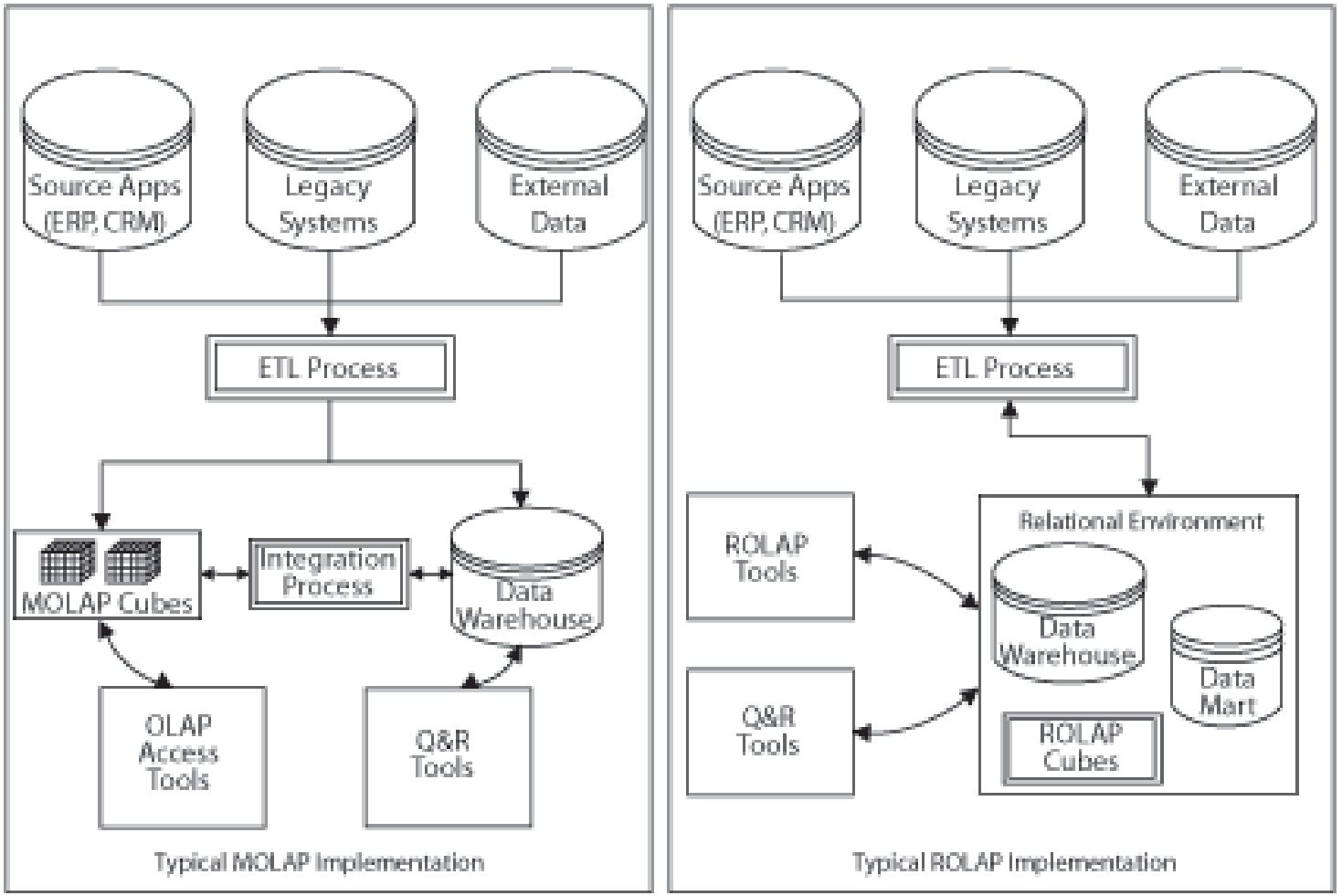
ROLAP – Relacional OLAP Server

- Estratégia pela qual são usados os próprios sistemas de banco de dados relacionais, com as tabelas sendo implementadas como estruturas relacionais clássicas.
- Oferece todas as vantagens de um SGBDR, porém exige um projeto cuidadoso do ponto de vista de desempenho, em que o excesso de tabelas normalizadas poderá comprometer a performance das buscas

Banco de Dados Relacional

Pros	Contras
Ideal para grande volume de dados	SQL não é otimizado para queries complexas
Tecnologia otimizada e aprovada	Determinar um esquema para armazenamento ótimo é mais importante e difícil.

MOLAP x ROLAP



ITnerante

Questão 10. DPE-SP - Agente de Defensoria - Administrador de Banco de Dados - 2010

A tecnologia OLAP feita em banco de dados relacionais que, por utilizar a estrutura relacional, possui a vantagem de não restringir o volume de armazenamento de dados é simulada pela arquitetura

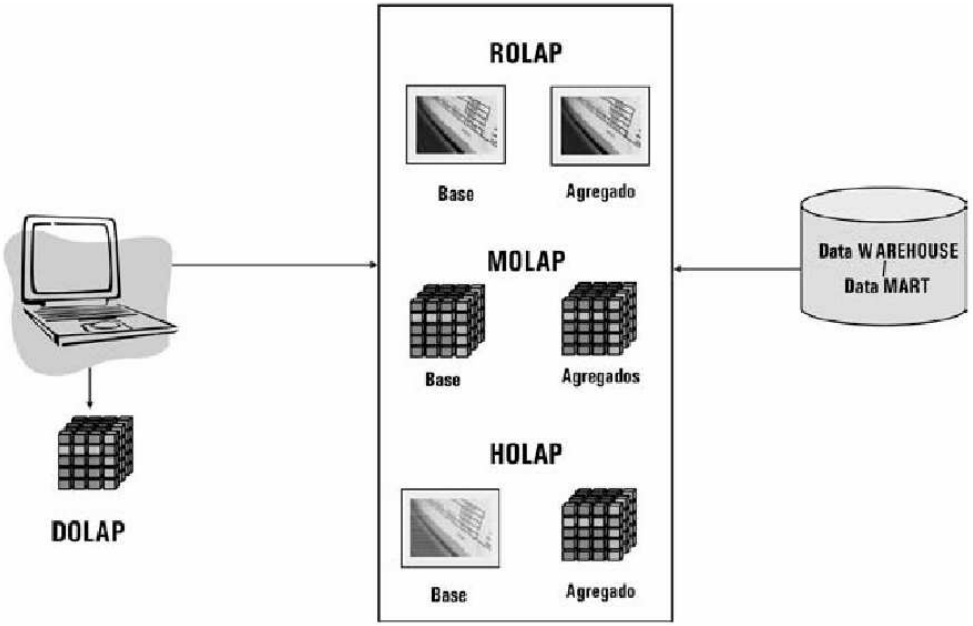
(a) HOLAP.

(b) ROLAP.

(c) DOLAP.

(d) WOLAP.

(e) MOLAP.



HOLAP E DOLAP

HOLAP - Hybrid OLAP Server

- Representa uma abordagem de uso misto das duas estratégias anteriores, em que:
  - As estruturas relacionais são normalmente utilizadas para os dados de menor granularidades e
  - As estruturas dimensionais nativas são dedicadas ao armazenamento de agregados (maior granularidade)

DOLAP – Desktop Olap Server

- Representa uma abordagem na qual estruturas dimensionais ou relacionais
  - Transferidas do DW/DM para as estações cliente
  - São armazenadas com o objetivo de facilitar o desempenho de certas análises
  - Minimizando o tráfego de informações entre o ambiente cliente e o ambiente servidor

Resumindo: Arquiteturas OLAP

- Classificadas em cinco tipos a seguir:
  - MOLAP (Multidimensional On Line Analytical processing);
  - ROLAP (Relational On Line Processing);
  - HOLAP (Hybrid On Line Analytical Processing);
  - DOLAP (Desktop On Line Analytical Processing);
  - WOLAP (Web On Line Analytical Processing).

ITnerante

TIMASTERS

OLAP

CESPE UnB
UNIVERSIDADE DE BRASÍLIA

FUNDAÇÃO CESGRANRIO

Fundação Carlos Chagas

ESAF
Escola de Administração Fazendária

FGV

Prof. Thiago Cavalcanti

Avaliação de produtos OLAP

12 regras de avaliação (Ted Codd)

1. Visão conceitual multidimensional
2. Transparência
3. Acessibilidade
4. Desempenho consistente na geração de relatórios
5. Arquitetura cliente-servidor
6. Dimensionalidade genérica
7. Manuseio dinâmico de matriz esparsa
8. Suporte a multiusuários
9. Operações irrestritas de cruzamento de dimensões
10. Manipulação de dados intuitiva
11. Relatório flexível
12. Dimensões e agregação de níveis ilimitados

Questão 11. ESAF - 2013 - DNIT - Analista Administrativo - Tecnologia da Informação

São regras de avaliação de produtos OLAP:
(a) Transferência ao usuário. Desempenho consistente na geração de relatórios. Dimensionalidade cumulativa. Operações irrestritas com dimensões cruzadas.

(b) Visão conceitual multidimensional para restringir consultas. Transparência ao usuário. Dimensionalidade genérica. Manipulação dedutiva dos dados.

(c) Visão conceitual multidimensional para formular consultas. Desempenho consistente na geração de relatórios. Dimensionalidade genérica. Manipulação intuitiva dos dados.

(d) Visão conceitual multidimensional para formular consultas. Dimensionalidade genérica. Manipulação segmentada dos dados. Operações irrestritas com dimensões alternadas.

(e) Extensão conceitual dos dados. Transparência ao dispositivo de acesso. Manipulação intuitiva dos dados. Operações irrestritas com indicações cruzadas.

Questão 12. CMV - 2010

55- Ferramentas de processamento analítico on-line (OLAP)

(a) funcionam sobre dados multidimensionais, caracterizados por atributos de dimensão e atributos de medida.

(b) funcionam sobre dados unidirecionais, caracterizados por atributos de medida e atributos de qualidade.

(c) funcionam sobre dados multidimensionais, caracterizados por atributos de dispersão e atributos de mediação.

(d) desconsideram dados multidimensionais.

(e) transformam dados unidimensionais em dados analíticos, caracterizando dimensão e medidas por atributos equivalentes.

Questão 13. CESPE - 2013 - MPOG - Tecnologia da Informação

Julgue os itens que se seguem, acerca das ferramentas ETL (extract transform load) e OLAP (on-line analytical processing).

[118] OLAP é uma tecnologia utilizada para organizar grandes bancos de dados e fornece, para organizações, um método com alta flexibilidade e desempenho para acessar, visualizar e analisar dados corporativos. Os dados podem ser organizados em uma hierarquia que define diferentes níveis de detalhe, na qual o usuário pode navegar para cima (drill up) ou para baixo (drill down) entre níveis.

Questão 14. CESPE - 2009 - DETRAN-DF - Analista - Análise de Sistemas

Acerca do desenvolvimento de aplicações e da arquitetura OLAP, julgue os itens a seguir.

[99] OLAP pode ser definido como o processo interativo de criar, gerenciar, analisar e gerar relatórios acerca de dados e deles exige algum tipo de agregação. Em bancos de dados multidimensionais (MOLAP), drill down significa ir de um nível mais baixo de agregação até um nível mais alto.

Questão 15. CESPE - 2011 - MEC - Gerente de Projetos

Com relação a modelagem de dados e OLAP, julgue os próximos itens.

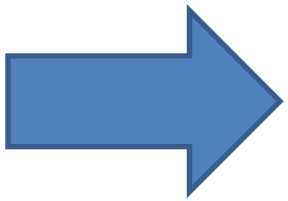
[125] As crosstabs, ou tabulações cruzadas, podem apresentar várias dimensões, em que são consideradas variáveis independentes, e a interseção entre as células da tabela contém valores de variáveis dependentes correspondentes a elas.

[126] Um drill down corresponde a ir de um nível mais baixo para um nível mais alto de agregação, ao passo que um drill up permite uma navegação pelas hierarquias em direção contrária.

[127] O banco de dados de um MOLAP possui um SGDB multidimensional, ou seja, permite armazenamento de dados nas células de um array multidimensional.

Crosstabs

Sample #	Gender	Handedness
1	Female	Right-handed
2	Male	Left-handed
3	Female	Right-handed
4	Male	Right-handed
5	Male	Left-handed
6	Male	Right-handed
7	Female	Right-handed
8	Female	Left-handed
9	Male	Right-handed
10	Female	Right-handed



	Left-handed	Right-handed	Total
Males	2	3	5
Females	1	4	5
Total	3	7	10

Questão 16. CESPE - 2011 - MEC - Gerente de Projetos

Com relação a modelagem de dados e OLAP, julgue os próximos itens.

[128] A agregação de dados em bancos SQL é necessária quando se faz qualquer tipo de processamento analítico, o que pode demandar um número muito grande de agrupamentos a serem considerados; entretanto uma consulta individual resulta em apenas uma tabela.

[129] Pivoteamento ou rotação é uma técnica para alterar uma hierarquia dimensional para outra em um cubo de dados.

[130] As projeções de uso de um datawarehouse que dá suporte a um OLAP são obtidas apenas após a construção dos dados no datawarehouse para que os caminhos de acesso e o armazenamento sejam sempre os mesmos, independentemente da demanda.

Questão 17. CESPE - 2012 - TCE-ES - Auditor de Controle Externo - Tecnologia da Informação

Acerca de data warehousing e OLAP, julgue os itens seguintes.

[135] Por meio da técnica denominada slice and dice, realiza-se a mudança de uma hierarquia dimensional para outra em um cubo de dados.

Questão 18. CESGRANRIO - 2010 - ELETROBRÁS - Analista de Sistemas - FUNCIONAL SAP-ERP

63 Analise as afirmações sobre as operações realizadas em cubos OLAP a seguir.

I - A operação pivot permite ao usuário alternar as linhas e colunas em que os valores visualizados serão recalculados.

II - A operação slice é caracterizada pela seleção de determinado membro de uma dimensão, a fim de analisar as demais informações do cubo sob tal perspectiva.

III - A operação dice corresponde à seleção específica de membros de duas ou mais dimensões.

Está correto o que se afirma em

(A) III, apenas.

(B) I e II, apenas.

(C) I e III, apenas.

(D) II e III, apenas.

(E) I, II e III.

Questão 19. TRT - 24ª REGIÃO (MS) - Analista Judiciário - Tecnologia da Informação - 2011

Quanto à perspectiva de visualização de um cubo, dentre as operações básicas OLAP,

(a) *slice* é a operação que corta o cubo, mas mantém a mesma perspectiva de visualização dos dados.

(b) *slice* é a mudança de perspectiva da visualização dos dados.

(c) *dice* é a operação que corta o cubo, mas mantém a mesma perspectiva de visualização dos dados.

(d) *slice* ocorre quando o usuário aumenta o nível de detalhe da informação, diminuindo o nível de granularidade.

(e) *dice* ocorre quando o usuário aumenta o nível de detalhe da informação, diminuindo o nível de granularidade.

Questão 20. TRT - 14ª Região (RO e AC) - Técnico Judiciário - Tecnologia da Informação - 2011

O usuário pode utilizar as ferramentas para navegar entre diferentes níveis de granularidade de um cubo de dados, aumentando ou diminuindo o nível de detalhamento dos dados, através de processos denominados *Drill up e Drill down*.

Trata-se de ferramentas aplicadas, tipicamente, em

(a) *OLAP*

(b) data *warehouse*

(c) data *mining*

(d) *OLAP e data warehouse.*

(e) *data warehouse e data mining.*

Questão 21. FUNRIO - 2013 - MPOG - Analista de Tecnologia da Informação

Na modelagem dimensional de dados para Data Warehouse, existem dois tipos de tabelas, representando os fatos contendo os dados granulares e os pontos de entrada específicos chamados de dimensões que descrevem os fatos. A modelagem dimensional facilita as consultas com operações OLAP (de Processamento Analítico On Line, em Inglês). A operação OLAP que permite relacionar fatos diferentes através de dimensões compartilhadas é denominada

(a) pivoteamento.
(b) drill down.
(c) drill up.
(d) drill across.
(e) slice and dice.

Questão 22. MPE-RN - Analista de Tecnologia da Informação - Engenharia de Software - 2010

Redução do escopo dos dados em análise, além de mudar a ordem das dimensões, mudando desta forma a orientação segundo a qual os dados são visualizados. Trata-se de uma operação OLAP de

(a) Slice and Dice.

(b) Drill Throught.

(c) Pivot.

(d) Roll Up.

(e) Drill Across.

Questão 23. TRT - 9ª REGIÃO (PR) - Técnico Judiciário - Tecnologia da Informação - 2010

Quando o usuário passa da análise da dimensão tempo e passa a analisar a dimensão região, por exemplo, ele está executando a operação OLAP

(a) *drill throught*.

(b) *slice and dice*.

(c) *drill across*.

(d) *roll up*.

(e) *star*.

Questão 24. TRF - 5ª REGIÃO - Analista Judiciário - Tecnologia da Informação - 2008

As consultas no star schema de um data warehouse podem ser feitas em maior ou menor nível de detalhe. Assim uma consulta mais detalhada das informações denomina-se

(a) drill-down.

(b) data mart.

(c) data mining.

(d) roll-up.

(e) snowflake.

Questão 25. TRIBUNAL REGIONAL DO TRABALHO DA 19 REG - Analista Judiciário – TI - 2011

Q.40. Considere:

I. Mudança de perspectiva da visão – extração de um subcubo.

II. Corta o cubo mas mantém a mesma perspectiva de visualização dos dados.

I e II correspondem, respectivamente, às operações básicas OLAP

(a) pivot e drill-throught .

(b) slice e dice .

(c) slice e pivot.

(d) dice e slice.

(e) dice e drill-across

Questão 26. TRE-AP - Analista Judiciário - Análise de Sistemas - 2011

Q.47.Uma das operações básicas de OLAP que ocorre quando é aumentado o nível de detalhe da informação é:

(a) *slice* and *dice*.

(b) *drill across.*

(c) *roll up.*

(d) *drill throught.*

(e) *drill down.*

Questão 27. TRT - 4ª REGIÃO (RS) - Analista Judiciário - Tecnologia da Informação - 2011

Q.45.Utilizando uma base multidimensional, o usuário passou da análise de informações sob a ótica da dimensão tempo para a visão sob a dimensão regional. A operação OLAP aí realizada foi

(a) roll up.

(b) drill across.

(c) drill throught.

(d) slice and dice.

(e) star across.

Questão 28. TCE-SP - Agente da Fiscalização Financeira - Informática - Produção e BD - 2010

A mudança de uma hierarquia (orientação) dimensional para outra tem sua realização facilitada em um cubo de dados por meio de uma técnica chamada

(a) roteamento.

(b) pivoteamento.

(c) ROLAP.

(d) OLTP.

(e) MOLAP.

Questão 29. SEFAZ-SP - Agente Fiscal de Rendas - Tecnologia da Informação - Prova 3 - 2009

As variáveis dimensionais aplicadas em um MOLAP estão frequentemente relacionadas em hierarquias, que determinam meios para agregar dados das células a elas associados. Nesse contexto, os operadores do processador que permitem percorrer (para acesso e não para criação) as hierarquias do nível de agregação mais baixo para o mais alto executam a função

(a) *snow flake.*

(b) *roll back.*

(c) *drill down.*

(d) *rolap.*

(e) *drill up.*

Questão 30. DPE-SP - Agente de Defensoria - Administrador de Banco de Dados - 2010

Um usuário pode pular um nível intermediário dentro de uma mesma dimensão por meio da operação OLAP do tipo

(a) drill down.

(b) drill up.

(c) drill throught.

(d) drill across.

(e) dlice and dice.

Questão 31. MPE-RN - Analista de Tecnologia da Informação - Banco de Dados - 2010

A arquitetura HOLAP (Hybrid On-Line Analytical Processing), para aproveitar as vantagens de alta performance e de escalabilidade, combina as tecnologias

(a) ROLAP e OLTP.

(b) ROLAP e MOLAP.

(c) DOLAP e MOLAP.

(d) OLAP e DOLAP.

(e) OLAP e OLTP.

Questão 32. MPE-RN - Analista de Tecnologia da Informação - Banco de Dados - 2010

Uma aplicação OLAP, com os dados armazenados no modelo relacional e também com suas consultas processadas pelo gerenciador relacional, deverá ter sua arquitetura elaborada com o método

(a) OLTP.

(b) MOLAP.

(c) ROLAP.

(d) DOLAP.

(e) HOLAP.

Questão 33. - CETRO - 2013 - ANVISA - Analista Administrativo - Área 5

Assinale a alternativa que apresenta um recurso essencialmente OLAP.

(a) ROLLUP.

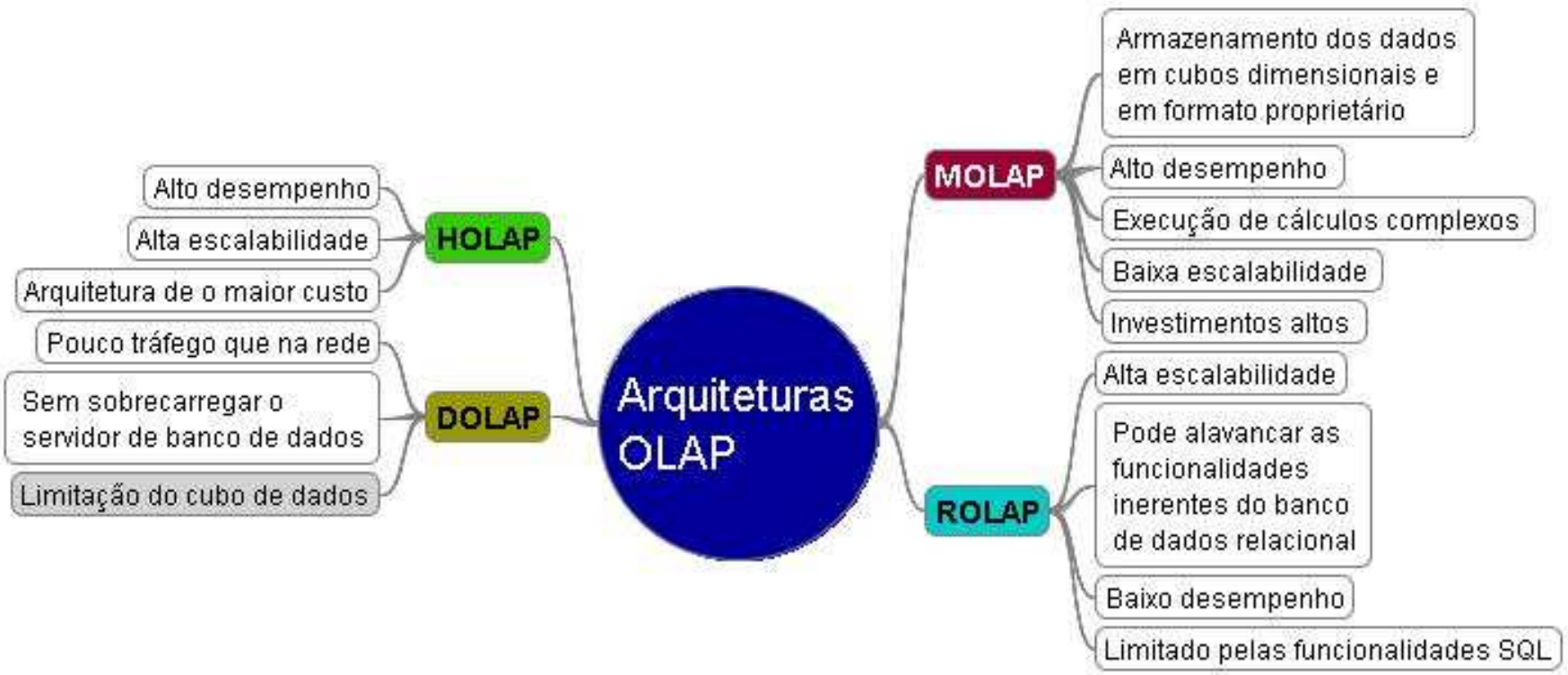
(b) GROUP BY.

(c) OLAPUP.

(d) HAVING.

(e) SELECT.

Mapa metal – Arquiteturas Olap



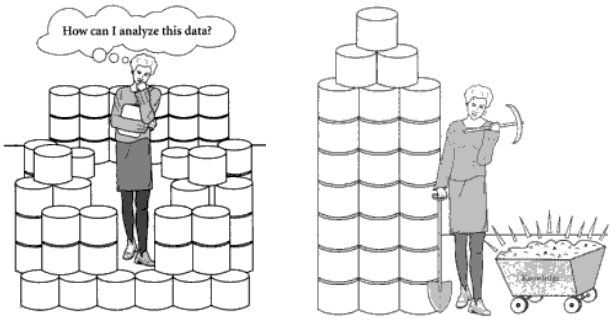
Gabarito – OLAP

1. D	12. A	23. A
2. C/C	13. C	24. A
3. C	14. E	25. D
4. A	15. C/C/C	26. E
5. B	16. E/C/E	27. C
6. B	17. E	28. B
7. E	18. E	29. E
8. B	19. A	30. C
9. E	20. A	31. B
10. B	21. D	32. C
11. C	22. A	33. A

ITnerante

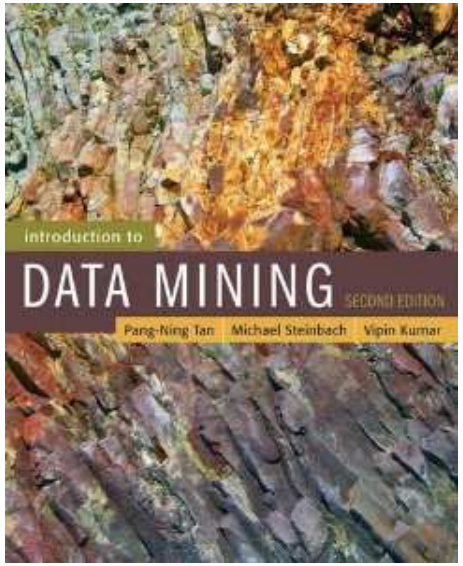
TIMASTERS

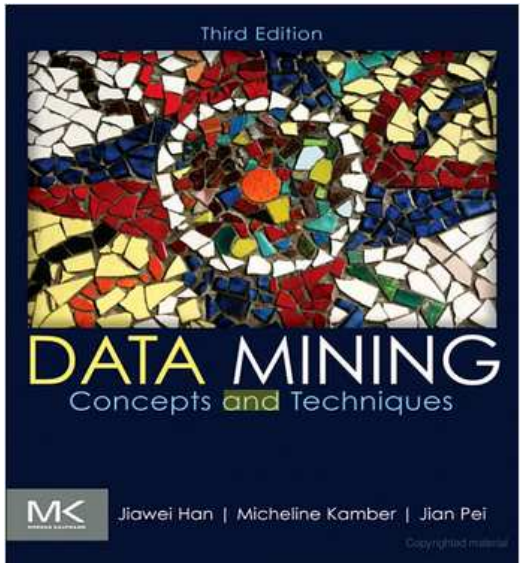
MINERAÇÃO DE DADOS

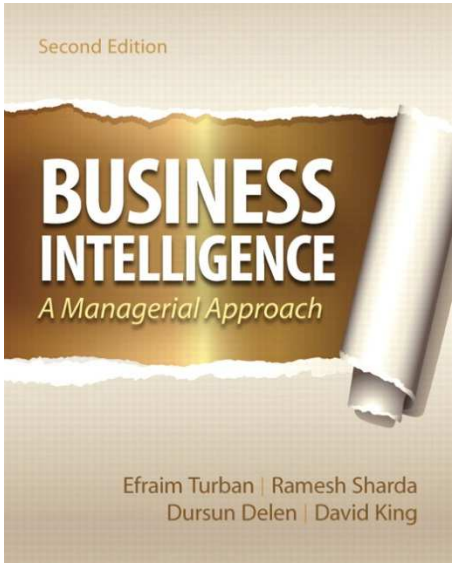


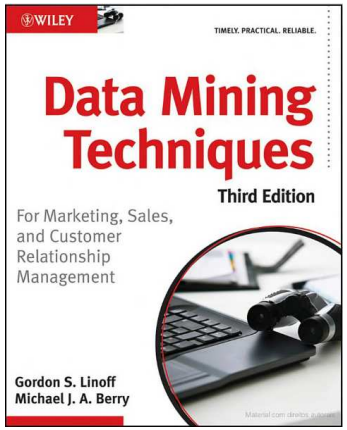
ITnerante

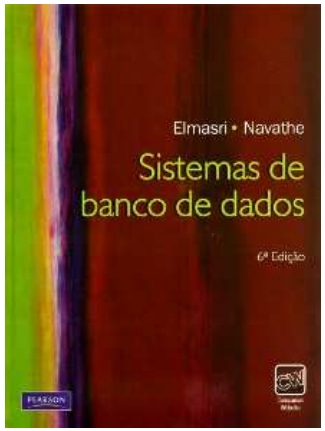
Bibliografia









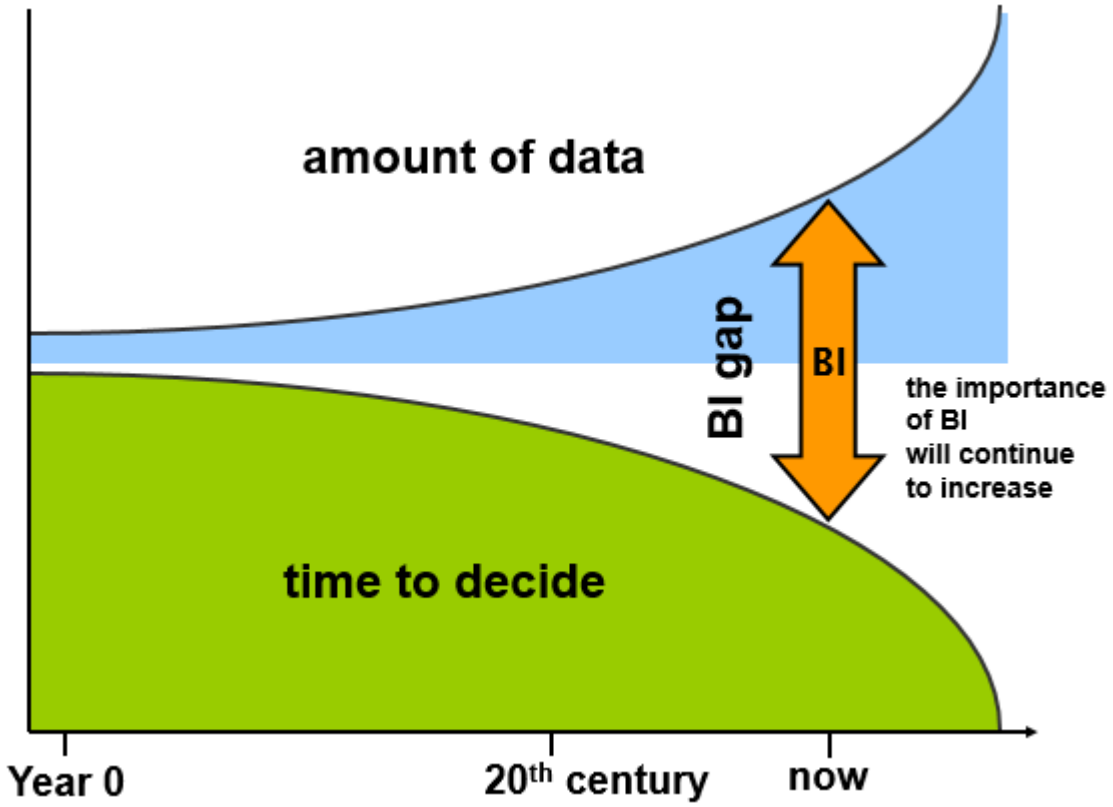


Universidade Federal do ABC

INPE

UFRGS

Motivação Inicial



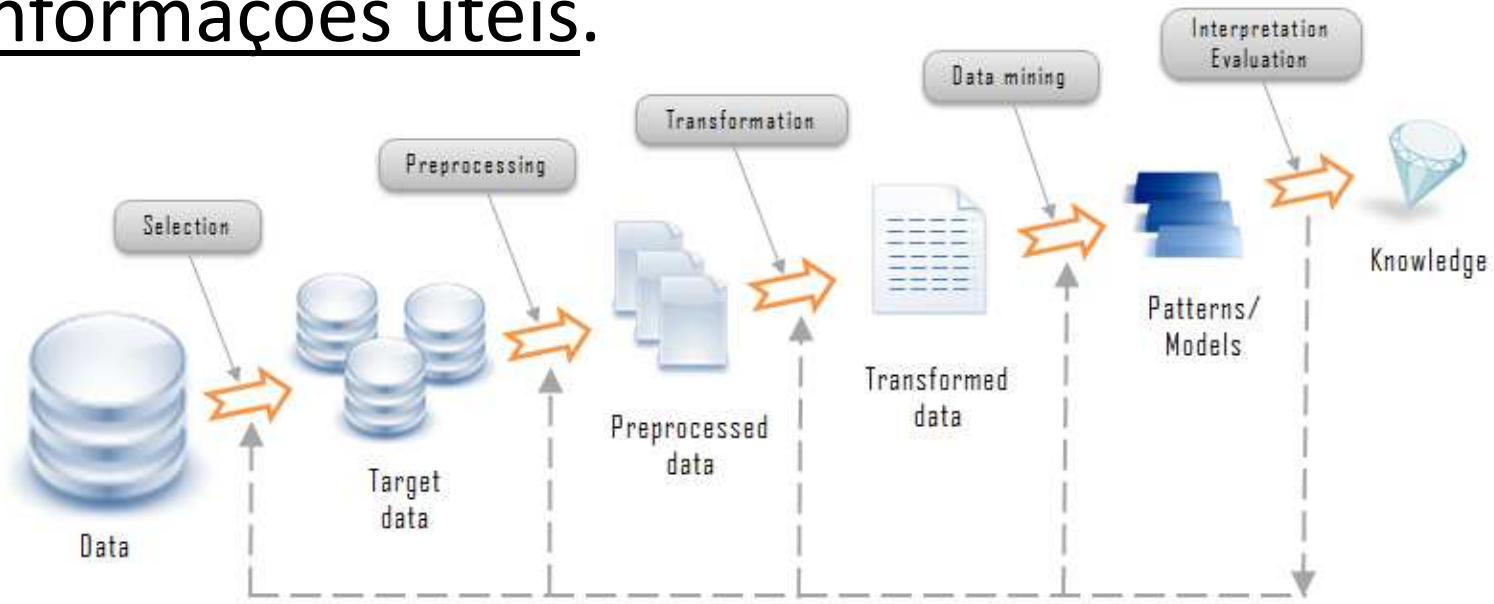
ITnerante

O que é Data Mining?

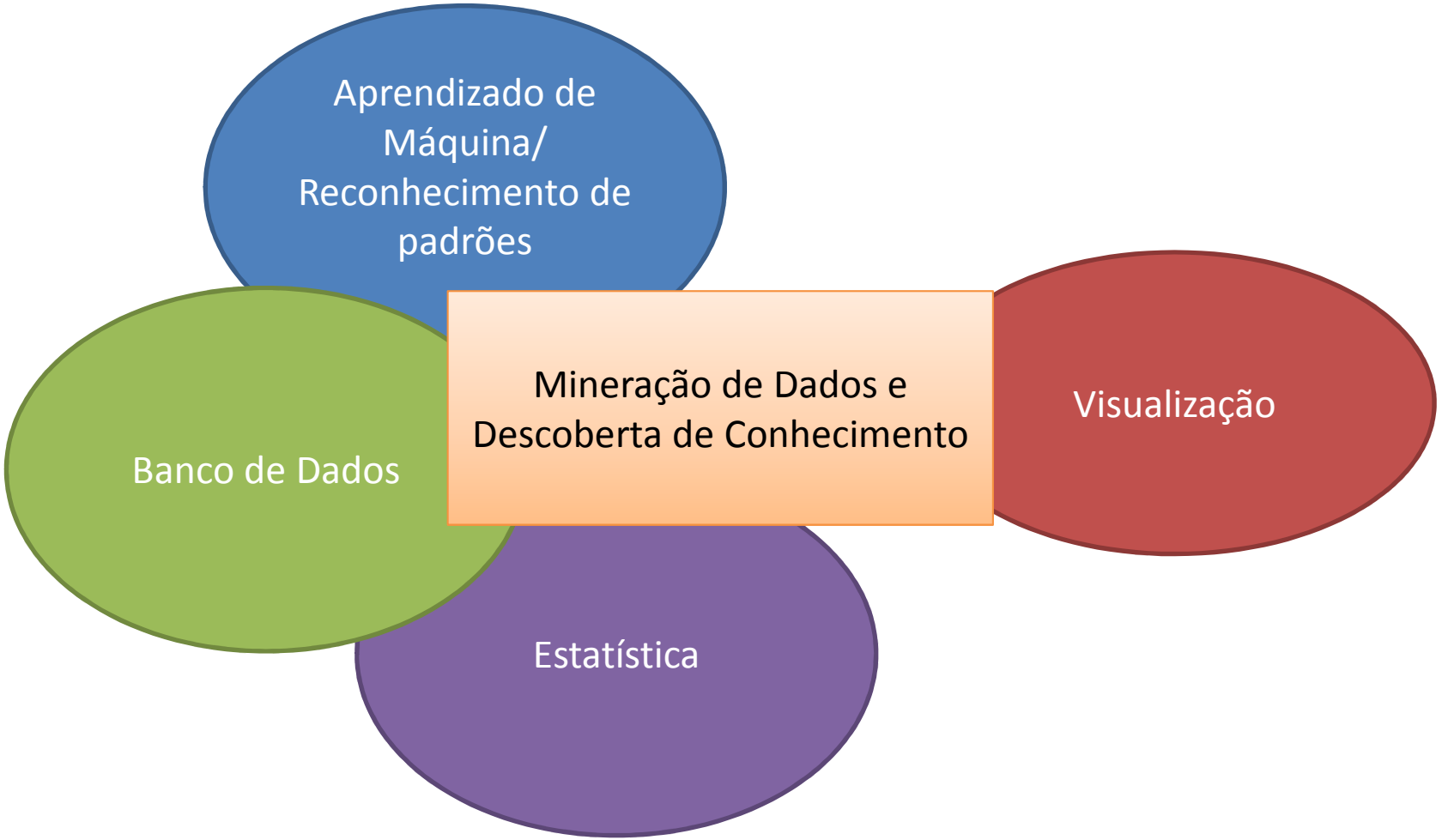
- Eduardo Gimenes: É o processo de **extrair informação válida**, previamente desconhecida e de máxima abrangência a partir de **grandes bases de dados**, usando-as para efetuar de **decisões** cruciais.
- Laudon&Laudon: **Análise** de grandes quantidades de **dados** a fim de encontrar **padrões e regras** que possam ser usadas para orientar a **tomada de decisões** e **prever o comportamento** futuro.

Definição: Mineração de dados

- Mineração de dados, ou data mining, é o processo de análise de conjuntos de dados que tem por objetivo a descoberta de padrões interessantes e que possam representar informações úteis.



Origens de Data Mining



Questão 01. FMP-RS - 2013 - MPE-AC - Analista - Tecnologia da Informação

Processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados é conhecido como

(a) *datawarehouse.*
(b) SGBD.
(c) mineração de dados *(data mining).*
(d) modelagem relacional de dados.
(e) mineração de textos *(text mining).*

Falácias de *Data Mining*

- ***Data Mining é automático:*** *é um processo, é iterativo,* requer supervisão.
- **Investimentos são recuperados rapidamente:** depende de muitos fatores!
- ***Software são intuitivos e simples****: é mais importante* conhecer os conceitos dos algoritmos e o negocio em si!
- ***Data Mining pode identificar problemas no negocio:*** *DM* pode encontrar padrões e fenômenos, identificar causa deve ser feito por especialistas.

Motivos que potencializam o uso

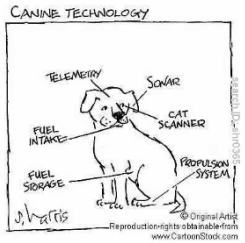
- O **volume de dados** disponível atualmente é enorme
- Os dados estão sendo **organizados**
- Os **recursos computacionais** estão cada vez mais potentes
- A **competição empresarial** exige técnicas mais modernas de decisão
- **Programas comerciais** de mineração de dados já podem ser adquiridos

Tarefas de Mineração de Dados

- A tarefa consiste na especificação **do que** queremos buscar nos dados
  - Tipo de regularidades ou categoria de padrões temos interesse em encontrar
  - Tipo de padrões poderiam nos surpreender (por exemplo, um gasto exagerado de um cliente de cartão de crédito, fora dos padrões usuais de seus gastos)

Classificação das tarefas

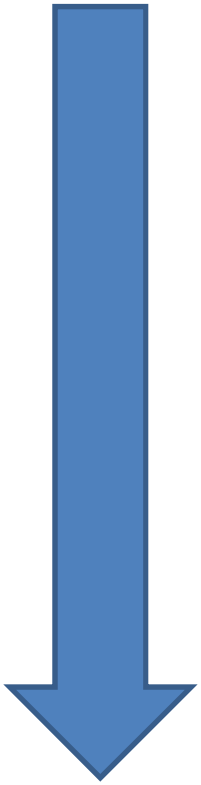
- Descritivas – caracterizam as **propriedades gerais** dos dados em um banco de dados
  - Achar **padrões** reconhecidos por seres humanos para descrever os dados



- Preditivas – essas tarefa **realiza uma inferências** sobre os dados atuais para fazer previsões sobre os mesmos
  - Usa variáveis para **prever valores futuros** ou desconhecidos de outras variáveis

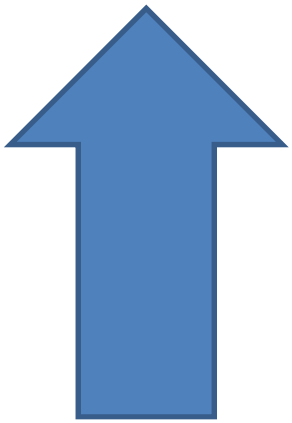
Classificação das tarefas

- Algumas dessas tarefas são melhor abordados de forma **top-down** chamado **teste de hipóteses**.
  - Em testes de hipóteses, um **comportamento** armazenado no banco de dados **passado** é utilizado para **verificar ou refutar** notações preconcebidas, ideias e palpites referentes às relações nos dados



Classificação das tarefas

- Outras tarefas são melhor abordadas de forma **bottom-up** chamado de **descoberta de conhecimento** (Knowledge discorvery).
  - Na descoberta de conhecimento, **sem suposições prévias** são feitas; os **dados** são autorizados a falar por si.



Segundo: Michael Berry, Gordon Linoff

- As **tarefas** adequadas para mineração de dados (não é limitado a essas):
  - **Classificação (Preditiva)**
  - Clustering (Descritiva)
  - Regra de Associação (Descritiva)
  - Regressão (Preditiva)
  - Detecção de desvios (Preditiva)

Técnicas de Mineração de Dados

- A técnica de mineração consiste na especificação de métodos que nos garantam **como** descobrir os padrões que nos interessam.
  - Dentre as principais técnicas utilizadas em mineração de dados, temos:
    - técnicas estatísticas
    - técnicas de aprendizado de máquina
    - técnicas baseadas em crescimento-poda-validação.

Características dos conjuntos de dados

- **Três características** aplicadas a muitos **conjuntos de dados** e que possuem um impacto significativo sobre **as técnicas de mineração** de dados: **dimensão**, **dispersão** e **resolução**.
  - A **dimensão** refere-se à quantidade de atributos de um conjunto de dados
  - A **resolução** está relacionada à granularidade dos dados
  - Um conjunto de dados é muito disperso quando para um atributo relevante, a maioria dos valores é NULL ou um valor padrão, e esse conceito está relacionado à **dispersão**

Questão 02 - ESAF - 2012 - CGU - Analista de Finanças e Controle - prova 3 - Auditoria e Fiscalização - Geral

São características gerais de conjuntos de dados:

(a) disposição, dispersão e renderização.

(b) dimensão, posicionamento e homogeneidade.

(c) compatibilidade, dispersão e interação.

(d) dimensão, dispersão e resolução.

(e) portabilidade, concentração e resolução.

Métodos para identificar padrões em dados

- Modelos simples
  - Consultas baseadas em SQL, OLAP
- Modelos intermediários
  - Regressão, árvore de decisão, agrupamento
- Modelos complexos
  - Redes neurais, outra indução de regras

Questão 03 - ESAF - 2012 - Receita Federal - Analista Tributário da RF - Prova 2 - Área Informática

Um *data mining* inteligente descobre informações em data *warehouses* onde consultas e relatórios não conseguem revela-las. Ferramentas de*data mining* encontram padrões em dados e podem até deduzir regras a partir deles. Os métodos usados para identificar padrões em dados são:

(a) modelos simples, modelos intermediários e modelos complexos.

(b) modelos simples, modelos físicos e modelos integrados.

(c) modelos híbridos, modelos *top-down* e modelos *bottom-up*.

(d) modelos lógicos, modelos físicos e modelos interativos.

(e) modelos básicos, modelos genéricos e modelos complementares.

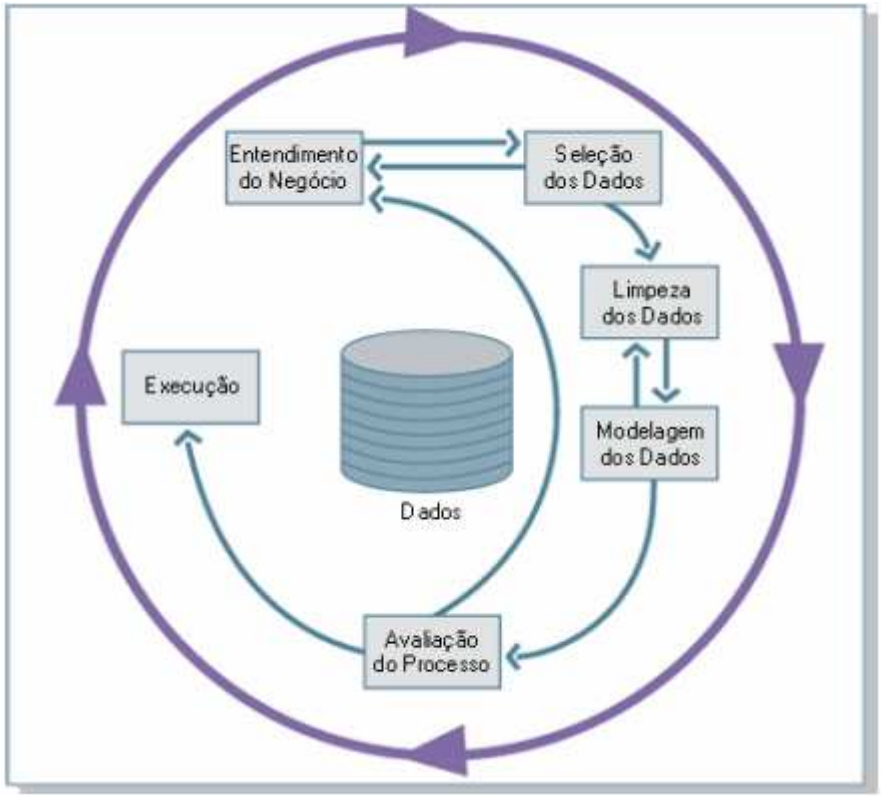
Desafios para Data Mining

- Escalabilidade - Scalability
- Alta dimensionalidade - Dimensionality
- Dados complexos e heterogêneos - Complex and Heterogeneous Data
- Qualidade dos dados - Data Quality
- Propriedade e distribuição de dados - Data Ownership and Distribution
- Preservação da privacidade - Privacy Preservation
- Dados em fluxo contínuo - Streaming Data

Questão 04 - ESAF - 2012 - CGU - Analista de Finanças e Controle - prova 3 - Auditoria e Fiscalização - Geral

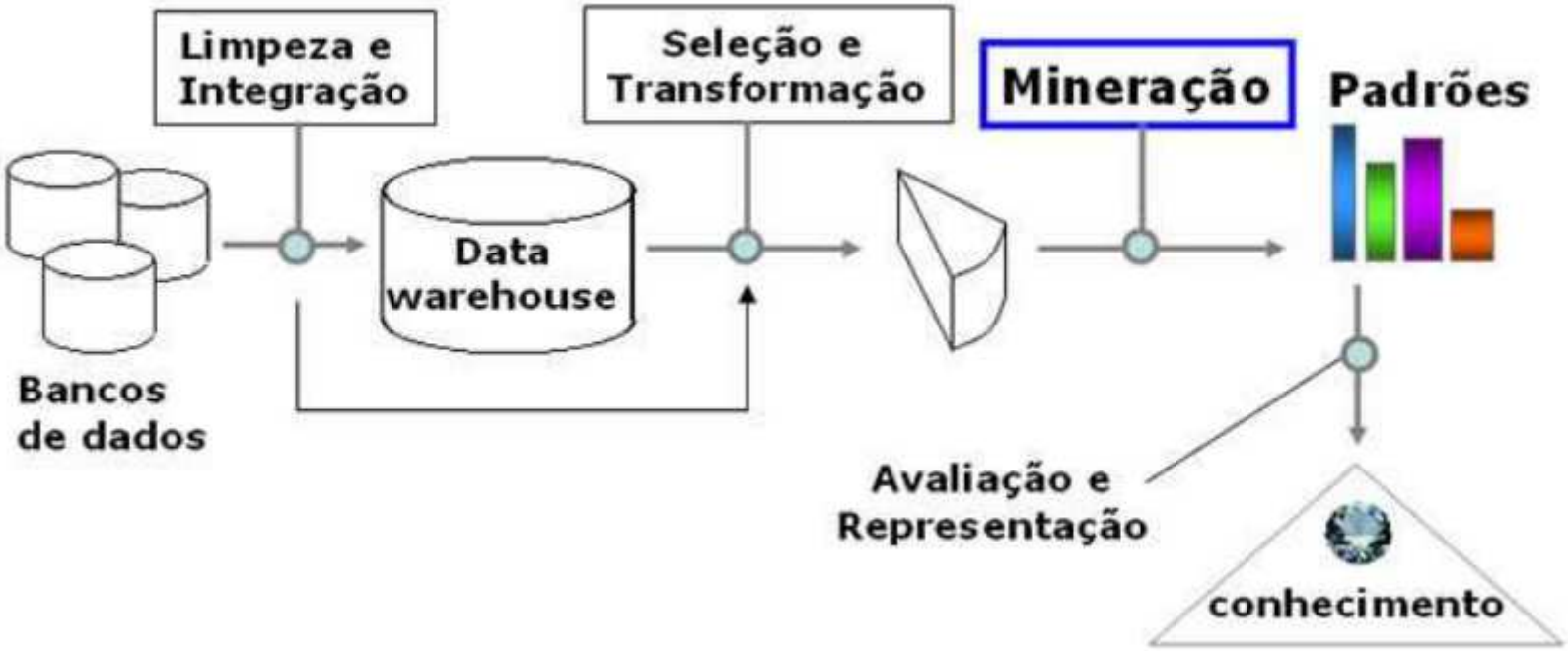
São aspectos motivadores da Mineração de Dados:

(a) Escalabilidade. Dimensionalidade moderada. Dados homogêneos. Propriedade e centralização dos dados.
(b) Extensibilidade. Alta paridade. Dados complexos e heterogêneos. Concorrência e distribuição dos dados.
(c) Escalabilidade. Alta dimensionalidade. Dados complexos e heterogêneos. Propriedade e distribuição de dados.
(d) Escalabilidade. Dimensionalidade variável. Dados compatíveis e acoplados. Adequação da distribuição de dados.
(e) Especialidade. Alta dimensionalidade de verificação. Dados complexos e complementares. Propriedade e consistência de dados.



PROCESSO DE MINERAÇÃO

Contextualizando (BI)

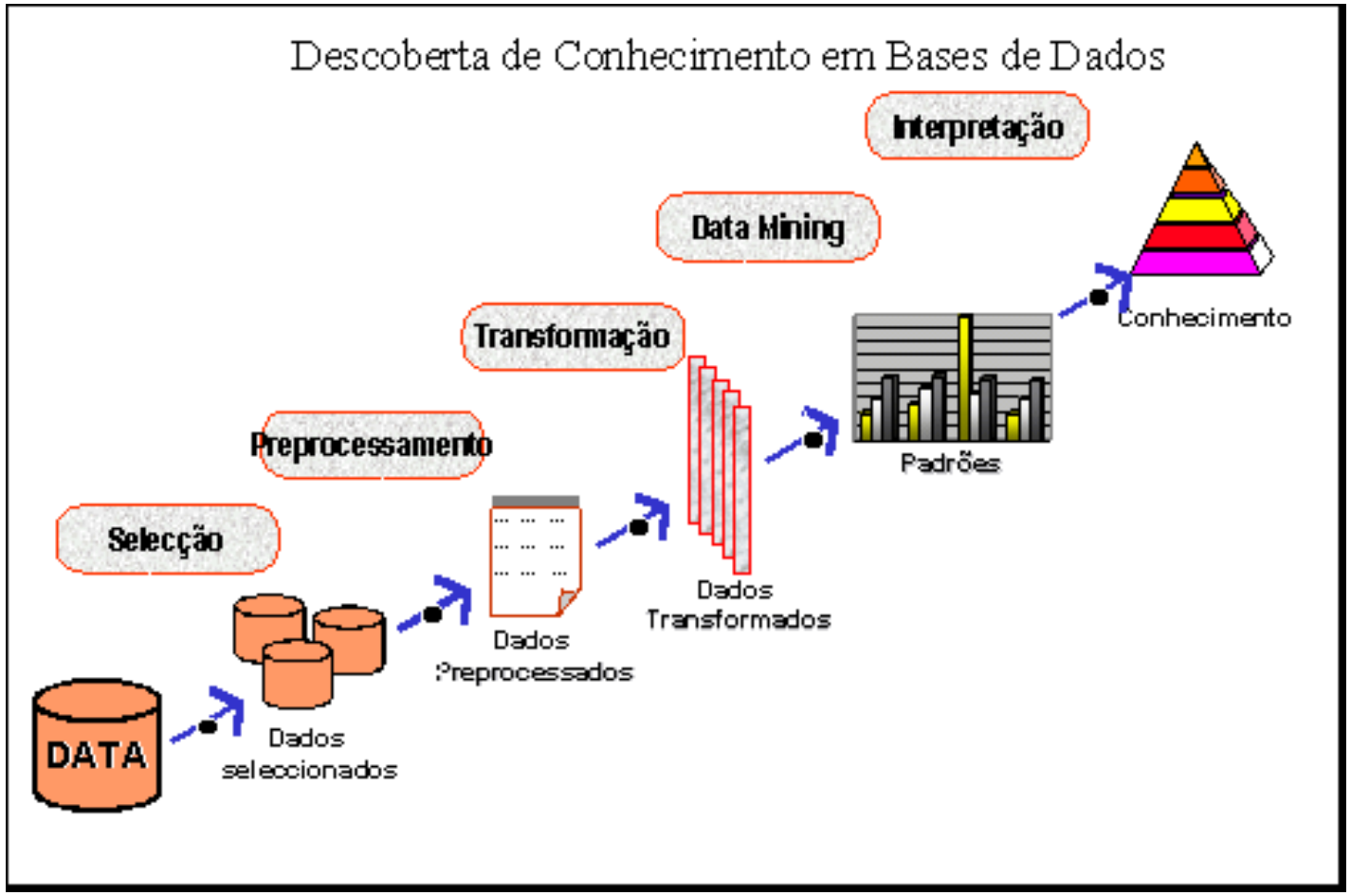


ITnerante

Knowledge Discovery in Databases - KDD

- KDD: Processo geral de **descoberta de conhecimentos úteis** previamente desconhecidos a partir de grandes bancos de dados
- Data Mining: Parte do processo de descoberta de conhecimentos em bancos de dados (KDD)

Descoberta de conhecimento (Fayad, 96)



ITnerante

Questão 05. INFRAERO - Analista Superior III - Analista de Sistemas - Administrador de BD -2011

Q.50. No âmbito da descoberta do conhecimento (KDD), a visão geral das etapas que constituem o processo KDD (Fayyad) e que são executadas de forma interativa e iterativa apresenta a seguinte sequência de etapas:

(a) seleção, pré-processamento, transformação, data mining e interpretação/avaliação.

(b) seleção, transformação, pré-processamento, interpretação/avaliação e data mining.

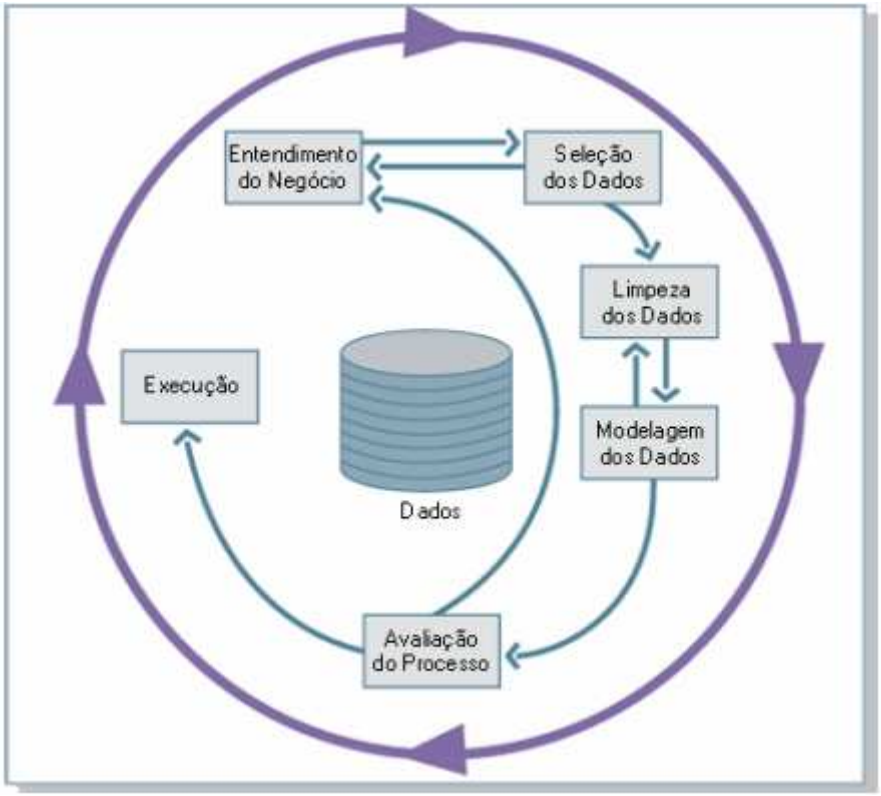
(c) data warehousing, star modeling, ETL, OLAP e data mining .

(d) ETL, data warehousing, pré-processamento, transformação e star modeling.

(e) OLAP, ETL, star modeling, data mining e interpretação/avaliação.

Fases da Mineração de Dados (CRISP-DM)

- Propõe uma **visão geral do ciclo de vida** de um **projeto** de mineração de dados



Entendimento do Negócio (Business Understanding)

- Foco no **entendimento do negócio** que visa obter conhecimento sobre os objetivos do negócio e seus requisitos.

Seleção dos Dados (Data Understanding)

- Consiste no **entendimento dos dados**, que visa à familiarização com o banco de dados pelo grupo de projeto, utilizando-se de conjuntos de dados "modelo".

Limpeza dos Dados (Data Preparation)

- A fase de **preparação de dados** consiste na preparação dos dados que visa a **limpeza**, **transformação**, **integração** e **formatação** dos dados da etapa anterior.

Modelagem dos Dados (Modeling)

- Fase que consiste na modelagem dos dados, a qual visa a **aplicação de técnicas** de modelagem sobre o conjunto de dados preparado na etapa anterior.
- Modela o problema em uma das técnicas baseadas em conceitos de:
  - Aprendizagem de máquina
  - Reconhecimento de padrões
  - Estatística

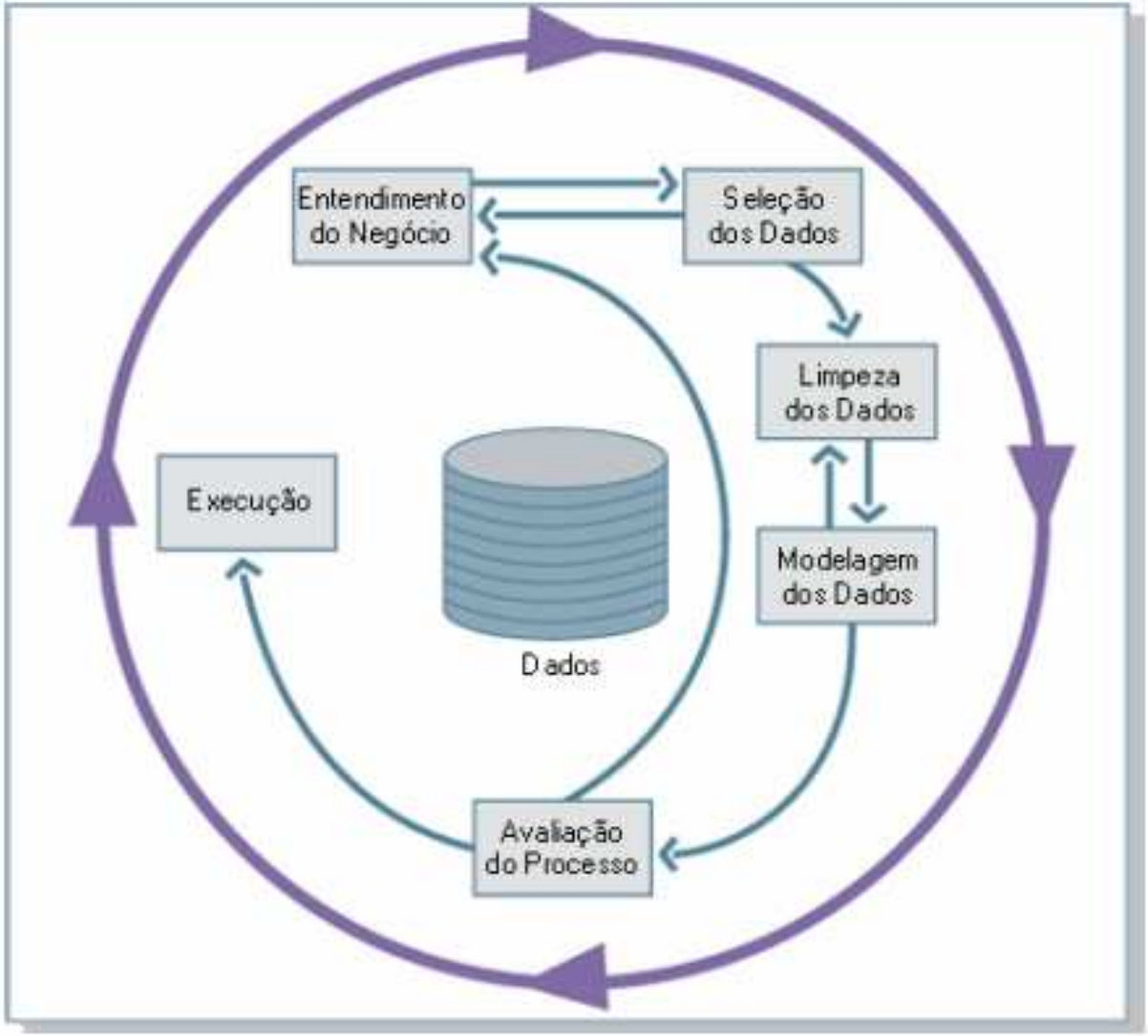
Avaliação do processo (Evaluation)

- Visa **garantir** que **o modelo** gerado atenda às **expectativas** da organização.
- Os **resultados** do processo de descoberta do conhecimento podem ser **mostrados** de diversas formas.

Execução (Deployment)

- Esta fase consiste na **definição das fases de implantação** do projeto de Mineração de Dados.

Fases da Mineração de Dados (CRISP-DM)



ITnerante

ITnerante

TIMASTERS

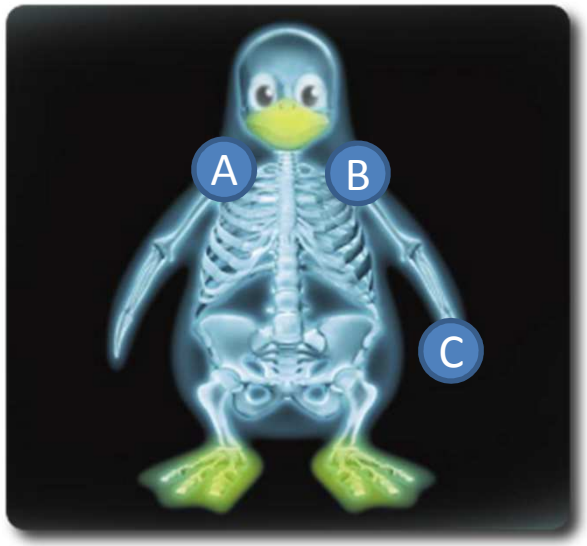
REGRAS DE ASSOCIAÇÃO

Prof. Thiago Cavalcanti

ITnerante

Regras de Associação

- Relacionam a presença de um conjunto de itens com outra faixa de valores de um outro conjunto de variáveis.



Análise de Regras de Associação

- Uma regra de associação é um padrão da forma X → Y , onde X e Y são conjuntos de valores

LME → LMD

conjunto-item

ITnerante

Como Avaliar os Padrões Interessantes ?

- **Suporte (%)**
  - Uma **medida objetiva** para avaliar o **interesse** de uma **regra de associação**
  - Representa **a porcentagem de transações (%)** de um banco de dados de transações **onde a regra se verifica**
- **Confiança**
  - Outra medida **objetiva** para **regras de associação**
  - Mede o **grau de certeza** de uma associação
  - Em termos estatísticos, trata-se simplesmente da **probabilidade condicional P (Y | X)**, isto é, a porcentagem de transações contendo os itens de X que também contém os itens de Y.

Questão 06 – EsFCEx- Escola da Formação Complementar do Exército - UFBA

[51] Análise o esquema seguinte e marque a alternativa correta que indica a regra de associação de compra cujas métricas de suporte e confiança indicam respectivamente os valores de 40% e 50%.

Considerando a teoria de mineração de dados e a tarefa de associação, a tabela abaixo ilustra algumas transações relacionadas a compras de supermercado. Os itens marcados com a letra 'X' indicam a ocorrência de compra do produto a cada transação.

Transação	Feijão	Leite	Arroz	Cerveja
1	X	X		X
2	X	X	X	
3			X	X
4	X	X	X	
5	X	X		X

A) {Arroz, Cerveja} → {Leite}
B) {Feijão, Leite} → {Arroz}
C) {Leite, Arroz} → {Feijão}
D) {Arroz} → {Cerveja, Leite}
E) {Cerveja} →{Feijão, Arroz}

Resolvendo

Transação	Feijão	Leite	Arroz	Cerveja
1	X	X		X
2	X	X	X	
3			X	X
4	X	X	X	
5	X	X		X

A) {Arroz, Cerveja} → {Leite}
B) {Feijão, Leite} → {Arroz}
C) {Leite, Arroz} → {Feijão}
D) {Arroz} → {Cerveja, Leite}
E) {Cerveja} →{Feijão, Arroz}

A) {Arroz, Cerveja} → {Leite} Suporte: 0% Confiança: 0%

B) {Feijão, Leite} → {Arroz} Suporte: 40% Confiança: 50%

C) {Leite, Arroz} → {Feijão} Suporte: 40% Confiança: 100%

D) {Arroz} → {Cerveja, Leite} Suporte: 0% Confiança: 0%

E) {Cerveja} → {Feijão, Arroz} Suporte: 0% Confiança: 0%

Suporte: % (LME U LMD)

Confiança: Suporte (LME U LMD)/ Suporte (LME)

Conjunto de itens grandes (itemset)

- Conjunto de itens que estejam acima dos limites estabelecidos para o **suporte** de uma regra de associação.
- Para cada conjunto de itens grandes, todas as regras que tenham um mínimo de confiança são gerados da seguinte forma:
  - Para cada conjunto grande X e Y C X, sendo Z = X – Y; então se o suporte (X)/suporte(Z)> confiança mínima, a regra Z => Y é uma regra válida.

A questão é como descobri todos os conjuntos de itens grandes?

- **Fechamento por baixo**
  - Um *itemset* grande também deve ser grande (ou seja, cada subconjunto de um *itemset* excede o suporte mínimo exigido)
- **Antimonotonicidade**
  - Um superconjunto de um *itemset* pequeno também é pequeno (implicando que ele não tem suporte suficiente)
  - Sendo assim quando se descobre um *itemset* pequeno, então qualquer extensão desse *itemset* será pequeno

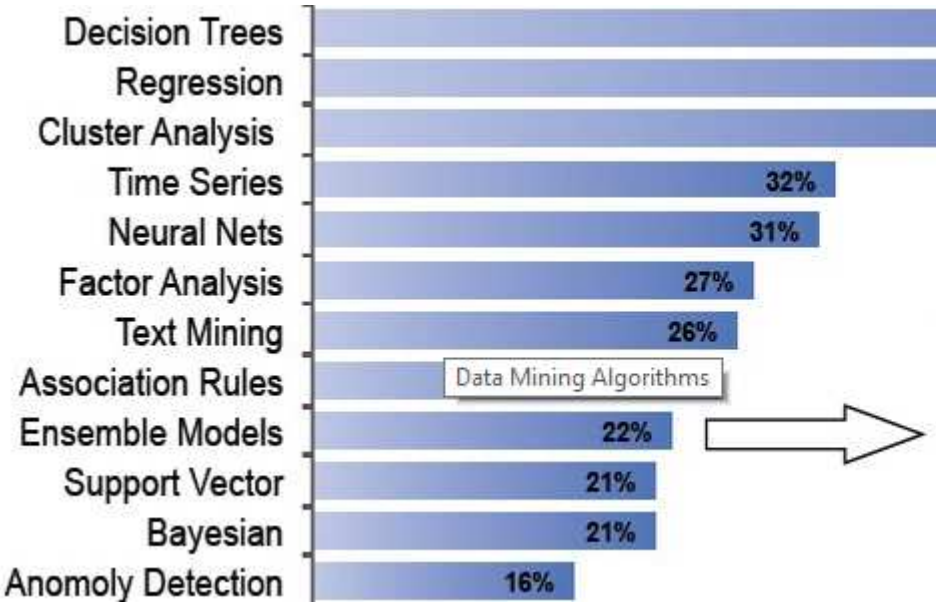
Questão 07. FUNRIO - 2013 - MPOG - Analista de Tecnologia da Informação

Qual o tipo de descoberta de conhecimento através de mineração de dados (do inglês “data mining”), em que se relaciona a presença de conjuntos de itens diversos, como por exemplo: “Quando uma mulher compra uma bolsa em uma loja, ela está propensa a comprar sapatos”?

(a) Hierarquias de classificação.
(b) Padrões sequenciais.
(c) Regras de associação.
(d) Séries temporais.
(e) Agrupamentos por similaridade.

Algoritmos para Data Mining

Regras de Associação



Consultants are more likely to use Ensemble Models

Corporate	Consultants	Academic	NGO / Gov't
21%	27%	20%	18%

Algoritmo a priori

Encontrar conjuntos de itens frequentes (grandes)

Entrada: banco de dados de m transações, D, e um suporte mínimo, *mins*, representado como uma fração de m.

Saída: itemsets frequentes, $L_1, L_2, ..., L_k$

Início /* etapas ou instruções são numeradas para aumentar a legibilidade */

1. Calcule suporte(i_j) = conta(i_j)/m para cada item individual, $i_1, i_2, ..., i_n$ fazendo a varredura do banco de dados uma vez e contando o número de transações em que o item i_j aparece (ou seja, conta(i_j));
2. O 1-itemset frequente candidato, C_1, será o itemset $i_1, i_2, ..., i_n$;
3. O subconjunto de itens contendo i_j de C_1 onde suporte(i_j) >= mins torna-se o 1-itemset frequente, L_1;
4. $k = 1$;
   termina = false;
   repita
1. $L_{k+1} = ; = ;$
2. Crie o $(k+1)$-itemset frequente candidato, C_{k+1}, combinando membros de L_k que têm $k-1$ itens em comum (isso forma os $(k+1)$--itemsets frequentes candidatos ao estender seletivamente os k-itemsets frequentes em um item);
3. Além disso, apenas considere como elementos de C_{k+1} aqueles $k+1$ itens tais que cada subconjunto de tamanho k apareça em L_k;
4. Faça a varredura do banco de dados uma vez e calcule o suporte para cada membro de C_{k+1}; se o suporte para um membro de C_{k+1} >= mins, então acrescente o membro em L_{k+1};
5. Se L_{k+1} for vazio, então termina = true, se não, $k = k + 1$;

   até que termina;

Fim;

Exemplificando

TID	A	B	C	D	E
T_1	1	1	1	0	0
T_2	1	1	1	1	1
T_3	1	0	1	1	0
T_4	1	0	1	1	1
T_5	1	1	1	1	0

Suporte: 40% (mínimo)

C_1

Itemset X	*supp(X)*
A	*?*
B	*?*
C	*?*
D	*?*
E	*?*

L_1

Itemset X	*supp(X)*
A	*100%*
B	*60%*
C	*100%*
D	*80%*
E	*40%*

C_2

Itemset X	supp(X)
A,B	?
A,C	?
A,D	?
A,E	?
B,C	?
B,D	?
B,E	?
C,D	?
C,E	?
D,E	?

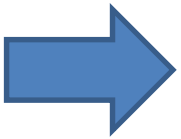
L_2

Itemset X	supp(X)
A,B	60%
A,C	100%
A,D	80%
A,E	40%
B,C	60%
B,D	40%
C,D	80%
C,E	40%
D,E	40%

Exemplificando

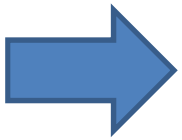
C_3

	Itemset X	$supp(X)$
join AB with AC	A,B,C	?
join AB with AD	A,B,D	?
join AB with AE	A,B,E	?
join AC with AD	A,C,D	?
join AC with AE	A,C,E	?
join AD with AE	A,D,E	?
join BC with BD	B,C,D	?
join CD with CE	C,D,E	?



C_3 after pruning

Itemset X	supp(X)
A,B,C	?
A,B,D	?
A,C,D	?
A,C,E	?
A,D,E	?
B,C,D	?
C,D,E	?

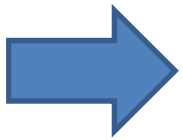


L_3

Itemset X	supp(X)
A,B,C	60%
A,B,D	40%
A,C,D	80%
A,C,E	40%
A,D,E	40%
B,C,D	40%
C,D,E	40%

C_4

	Itemset X	$supp(X)$
combine ABC with ABD	A,B,C,D	?
combine ACD with ACE	A,C,D,E	?



L_4

Itemset X	supp(X)
A,B,C,D	40%
A,C,D,E	40%

Algoritmo por Amostragem

- Objetiva selecionar um pequena amostra, que caiba na memória principal do banco de dados transacional, e determinar os conjuntos de itens frequentes daquela amostra.
  - Usar o algoritmo a priori com suporte diminuído.
  - Borda negativa
    - Importante porque determina o suporte para aqueles conjuntos de itens, garantido que nenhum conjunto de itens grande foi perdido na análise da amostra de dados.

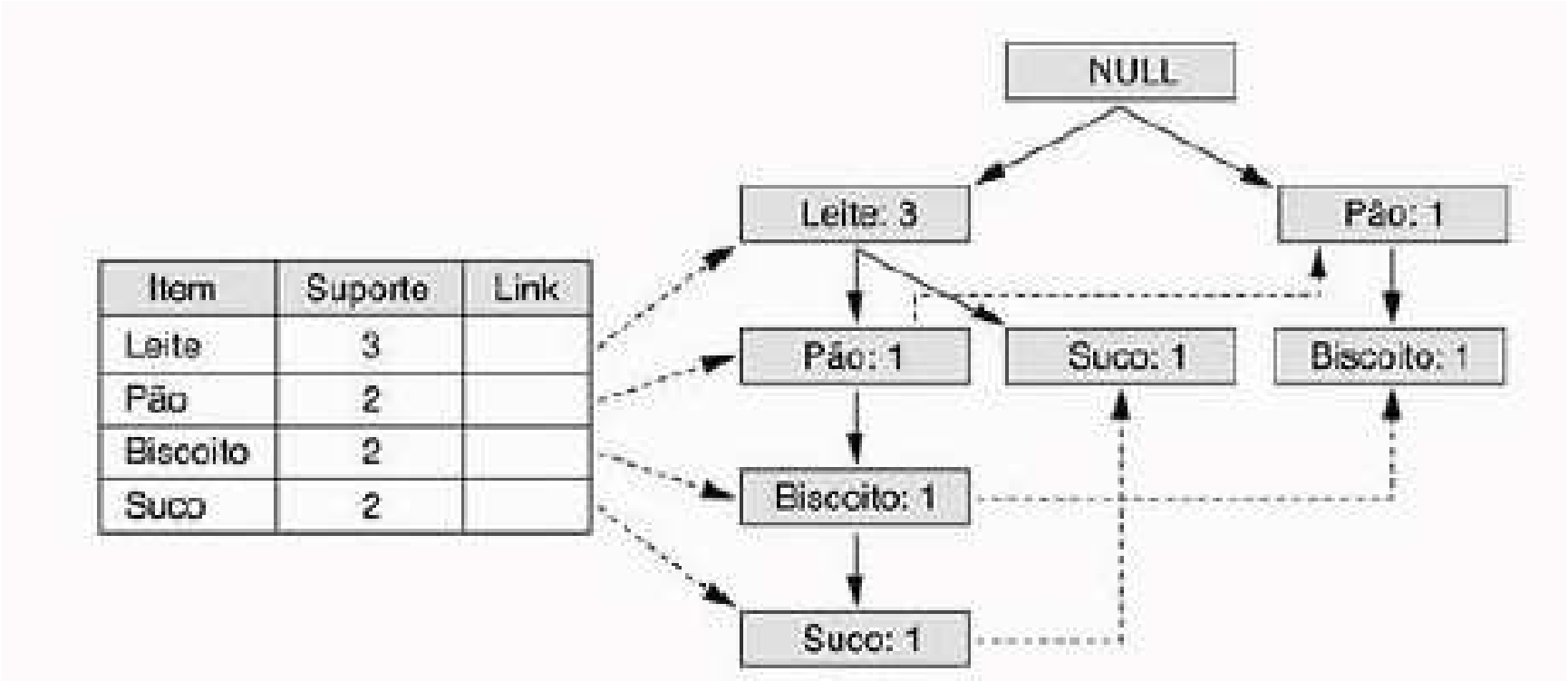
Árvore de padrão frequente (Árvore FP)

- É motivado pelo fato **de algoritmos apriori** poderem gerar e testar um número muito grande de itemsets candidatos.

$$\left(\frac{1000}{2}\right) = 499500$$

- O algoritmos de crescimento FP é uma técnica que elimina a geração de um grande número de itemsets candidatos.

Árvore de padrão frequente (Árvore FP)



Item	Suporte	Link
Leite	3	
Pão	2	
Biscoito	2	
Suco	2	

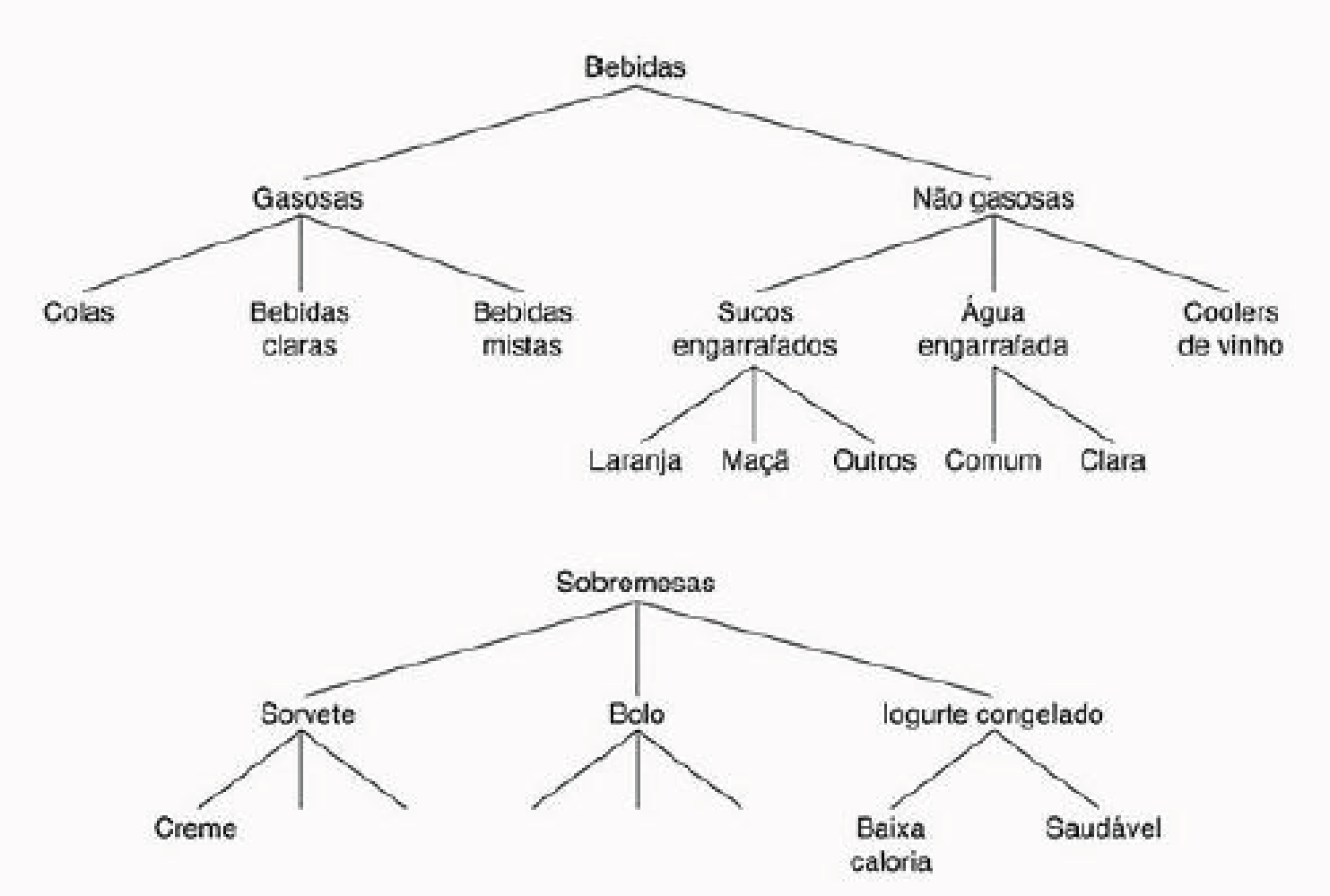
Algoritmos de Partição

- Se tivermos um banco de dados com um número **potencial pequeno de conjunto de itens grandes**, digamos, alguns poucos milhares, então o suporte para todos eles pode ser testado em uma passagem pela técnica de partição.
- A partição **divide o banco em subconjuntos** sem sobreposição; estes são considerados banco de dados separados, e todos os conjuntos de itens grandes para cada partição, chamados **conjuntos de itens frequentes locais**, são gerados em uma única leitura do banco.

Regras de associações entre hierarquias

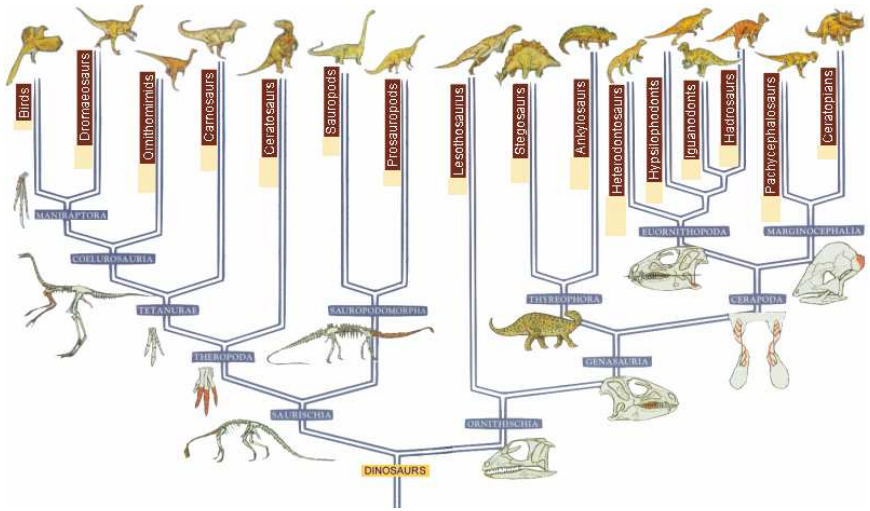
- Em geral, é possível dividir os itens em **hierarquias disjuntas** com base na natureza do domínio.
- Se a área de aplicação tiver uma classificação natural dos conjuntos de itens em hierarquias, descobrir associações dentro das hierarquias **não** tem qualquer interesse particular.
- O interesse específico **são associações entre hierarquias**.
  - Eles pode ocorrer entre agrupamentos de item em diferentes níveis.

Taxonomia de itens em um supermercado



Classificação

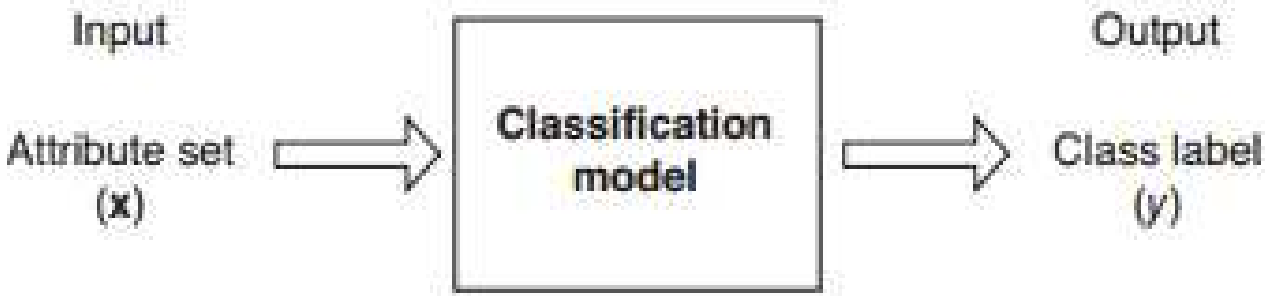
Conceitos Básicos



Classificação

- Uma das tarefas **mais comuns** dentro de mineração de dados
- Consiste em examinar as características de um objeto recém apresentados e atribuí-lo a um dos conjuntos predefinidos de classes
- A tarefa de classificação é caracterizada por **uma definição das classes(1)**, e conjunto **dados para aprendizado(2)** pré-classificados

Classificação segundo TAN



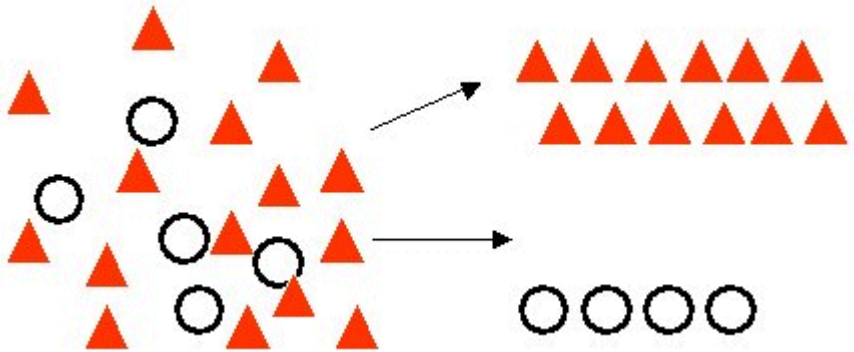
- É a tarefa de aprendizado de **uma função alvo** f que mapeia **cada atributo** de um **conjunto x** para **um rótulo** de classe predefinido y.

Questão 08. ESAF - 2012 - CGU - Analista de Finanças e Controle - prova 3 - Auditoria e Fiscalização - Geral

Classificação é
(a) a tarefa de atualizar uma função focal *f* que permeia cada conjunto de variáveis *x* para um dos blocos de classes *y* discretos.
(b) o mapeamento de uma função objetivo *f* à qual são atribuídos valores *x* fixados por categorias de rótulos de classes *z* pré-determinados.
(c) a função alvo *f* que mapeie cada classificação de atributos *x* para um dos eixos de classes *y* pré-determinados.
(d) a tarefa de aprender uma função alvo *f* que mapeie cada conjunto de atributos *x* para um dos rótulos de classes *y* pré-determinados.
(e) a tarefa de ordenar funções de mapeamento para cada categoria de atributos *x* para um dos rótulos de variáveis *y* controladas.

Classificação segundo Navathe

- É o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem **classes** ou **conceitos**
  - Tem o propósito de utilizar o modelo para **predizer a classe de objetos** que ainda não foram classificados
  - **Aprendizado supervisionado**



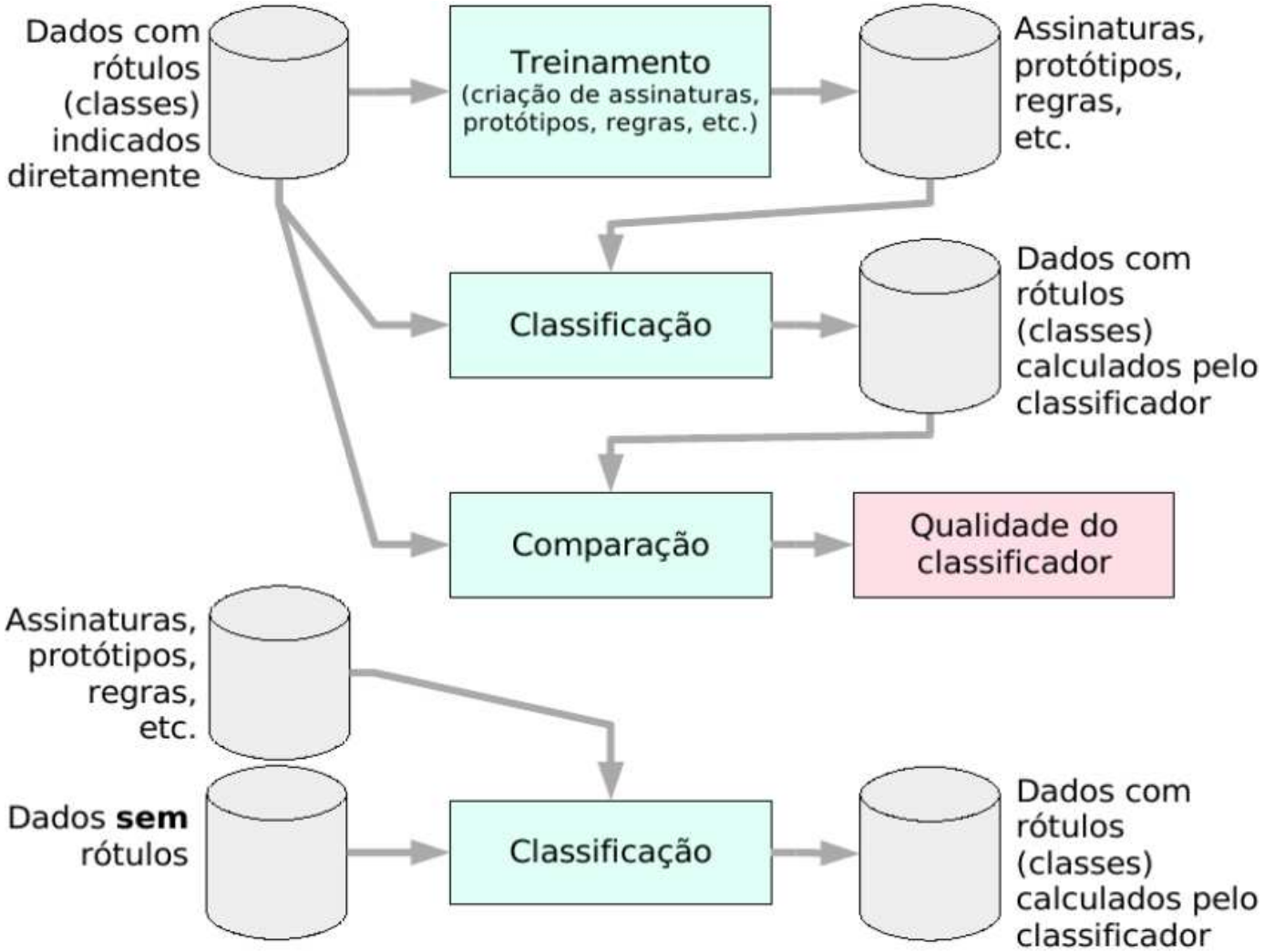
Classificação

- A objetivo é a construção de um modelo que possa ser aplicado a dados não classificados e classificá-los
  - Exemplos de tarefas de classificação que foram abordados através de técnicas de mineração de dados:
    - Classificação de pedido de crédito como baixo, médio ou alto risco
    - Escolher conteúdo a ser exibido em uma página Web
    - Determinar quais os números de telefone correspondem a máquinas de fax
    - Descobrir sinistros fraudulentos
    - Atribuir códigos da indústria e denominações de emprego com base nas descrições de texto livre

Classificação

- Em todos os exemplos, há **um número limitado de classes**, e espera-se ser capaz de atribuir qualquer registo em um ou outra.
- As **árvores de decisão** e técnicas semelhantes são bem adaptadas para a classificação.
- **Rede neural** e análise de links também são úteis para a classificação de certas circunstâncias

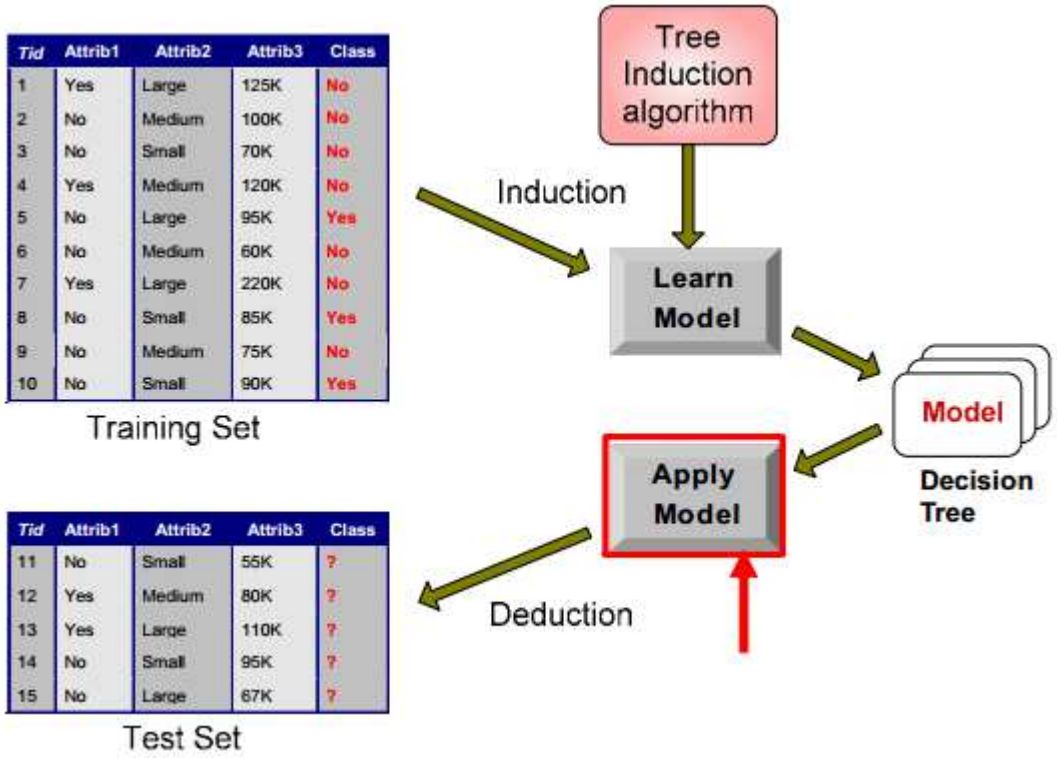
Classificação



ITnerante

Arvore de decisão

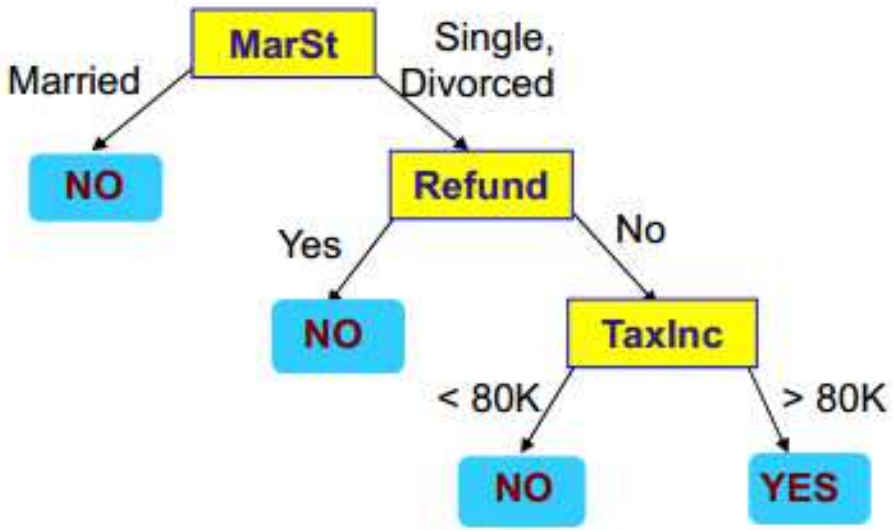
- Uma representação gráfica da **descrição** de cada classe ou das **regras** de classificação



ITnerante

Exemplo de Árvore de Decisão

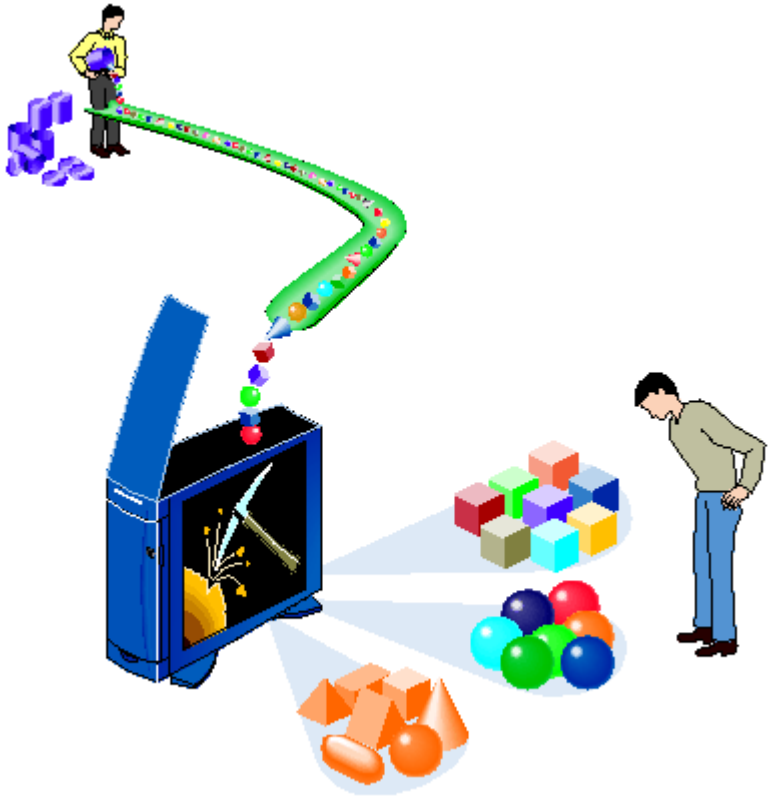
	categórico	categórico	contínuo	classe
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Pode haver mais de um árvore para o mesmo conjunto de dados

Classificação (Resumo)

- Tarefa: Dado um conjunto de exemplos pré-classificados, construir um modelo ou um classificador para classificar novas entradas.
- Aprendizado supervisionado
- Um classificador pode ser um conjunto de regras, uma árvore de decisões, uma rede neural, ...
- Algumas aplicações:
  - Aprovação de crédito, marketing direto, detecção de fraudes, diagnóstico médico ...



AGRUPAMENTO (CLUSTERING)

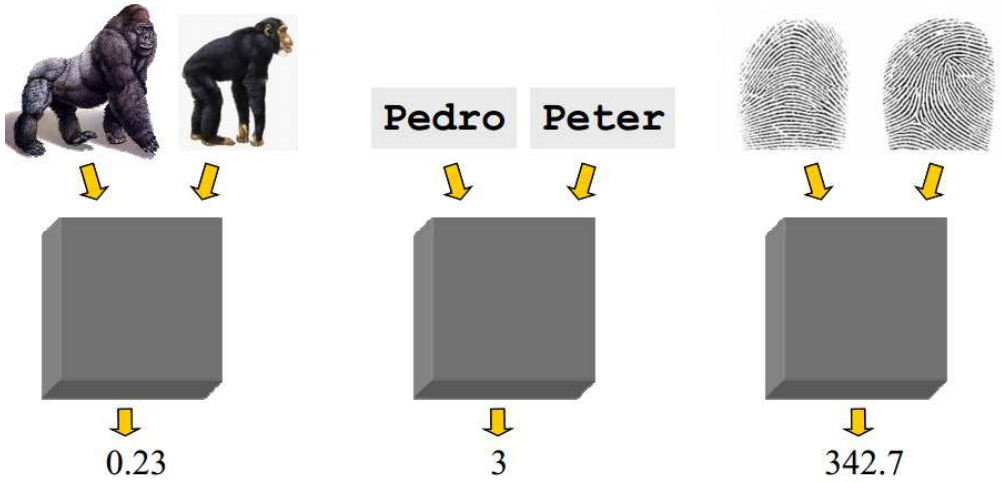
Prof. Thiago Cavalcanti

Análise de Clusters (Agrupamentos)

- Consiste em identificar **agrupamentos de objetos**, estes que identificam **uma classe**
- Trabalha sobre dados onde as etiquetas das classes **não estão definidas**.
- Conhecido também por **aprendizado não supervisionado** e, às vezes, chamado de classificação por estatísticos e de segmentação por pessoas de marketing

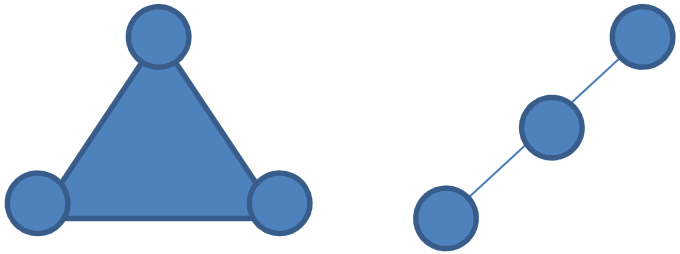
Distância (Definição)

- Medidas de Distância
  - Sejam O1 e O2 dois objetos de um universo de possíveis objetos. A distância (dissimilaridade) entre O1 e O2 é um número real denotado por D (O1,O2)



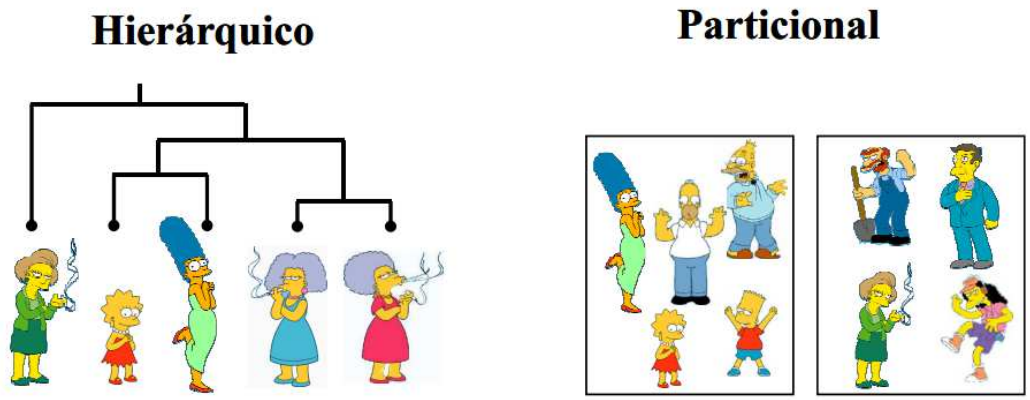
Propriedades de uma medida de distância

- Simetria
  - D (A,B) = D (B,A)
- Constância de auto-similaridade
  - D (A,A) = 0
- Positividade
  - D(A,B) = 0 ⇔ A = B
- Desigualdade Triangular
  - D (A,B) ≤ D(A,C) + D (B,C)



Tipos de Agrupamento

- Algoritmos Particionais:
  - Construir diversas partições e avaliá-las com algum critério
- Algoritmos Hierárquicos:
  - Criar uma decomposição hierárquica de um conjunto de objetos utilizando algum critério



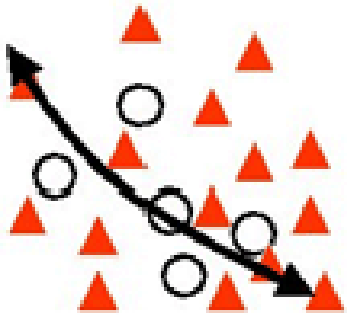
Questão 09 - CESPE - 2013 - MPU - Analista - Suporte e Infraestrutura

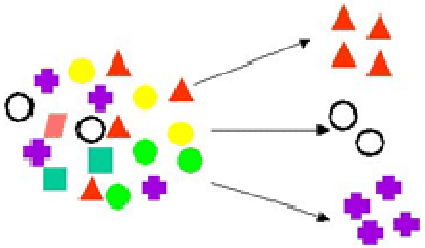
Julgue os próximos itens, acerca de sistemas de suporte à decisão.

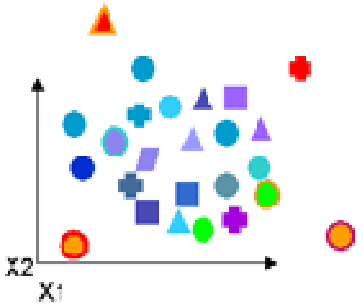
[100] Em se tratando de mineração de dados, a técnica de agrupamento (clustering) permite a descoberta de dados por faixa de valores, por meio do exame de alguns atributos das entidades envolvidas.

ITnerante

TIMASTERS







ABORDAGENS PARA OUTROS PROBLEMAS DE DATA MINING

ITnerante

Análise de Padrões Sequenciais

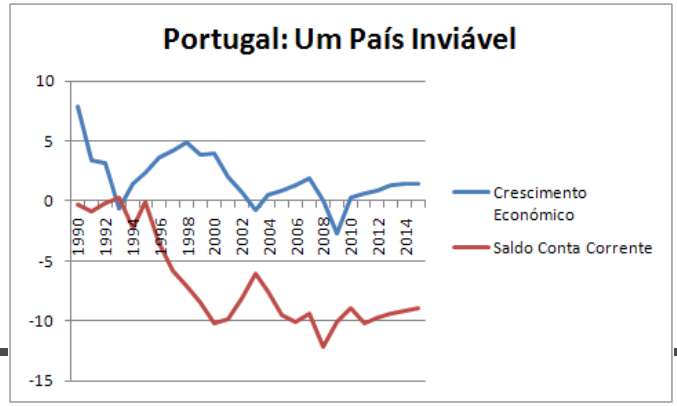
- Um padrão sequencial é uma expressão da forma < i1;.....; in >, onde cada i é um conjunto de itens.
  - A ordem em que estão alinhados estes conjuntos reflete a ordem cronológica em que aconteceram os fatos representados por estes conjuntos

Análise de Padrões em Séries Temporais

- Ex: O preço de fechamento uma ação ou de um fundo é um evento que ocorre a cada dia da semana para cada fundo ou ação.

  → Sequencias desse valores é uma serie temporal

- Séries temporais são sequencias de eventos; cada evento pode ser um tipo fixo dado uma transação.



ITnerante

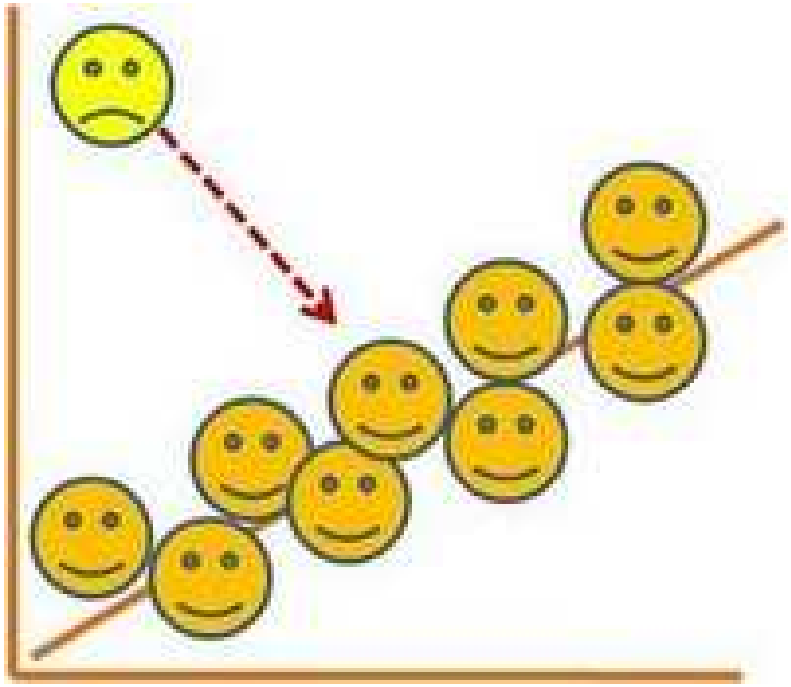
Predição

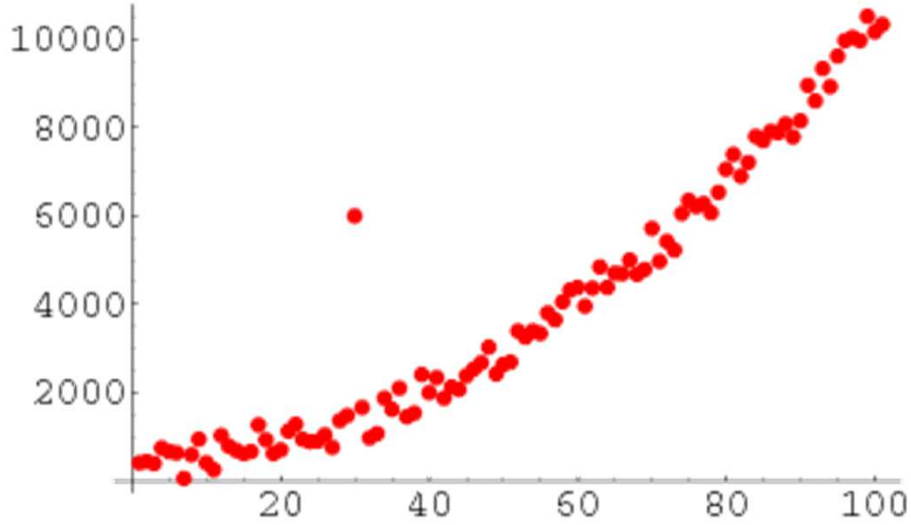
- Em algumas aplicações, o usuário está mais interessado em predizer alguns valores ausentes em seus dados, em vez de descobrir classes de objetos.
- Isto ocorre sobretudo quando os valores que faltam são numéricos.

Análise de Outliers

- Um banco de dados pode conter dados que não apresentam o comportamento geral da maioria.
  - Estes dados são denominados *outliers* (exceções)
  - Muitos métodos de mineração descartam estes *outliers* como sendo **ruído indesejado**
  - Entretanto, em algumas aplicações, tais como **detecção de fraudes**, estes eventos raros podem ser mais interessantes do que eventos que ocorrem regularmente.

Análise de Outliers





ITnerante

Regressão

- Regressão é uma aplicação especial da regra de classificação.
  - Se uma regra de classificação é considerada uma **função sobre variáveis** que as mapeia em uma classe destino, a regra é chamada **regressão**.
- Uma aplicação de regressão ocorre quando, em vez de mapear uma tupla de dados de uma relação para uma classe específica, o valor da variável é previsto baseado naquela tupla.

Regressão Linear

- Quando:
  - Y = f(x1, x2, ..., xn)
  - Um função f é linear no domínio das variáveis xi, o processo de derivar f de um dado conjunto de tuplas para <x1, x2, ... Xn, y> é chamado **regressão linear**.

X1	X2	X3	X4	...	XN
x1	x2	x3	x4	...	xn
z1	z2	z3	z4	...	zn

Questão 10. MPOG – APO 2008

27- São tarefas primárias da Mineração de Dados:

a) Classificação; Regressão; Clusterização.

b) Classificação; Realimentação; Complementação.

c) Codificação; Normalização; Clusterização.

d) Composição; Migração; Clusterização.

e) Compressão; Processamento; Associação.

Questão 11. MP DO ESTADO DO PARÁ – Cargo: ANALISTA DE SISTEMA – SUPORTE A BD

58. Considerando as etapas e tarefas de Mineração de Dados, é correto afirmar que

(A) a Mineração é a etapa essencial do processo; ela consiste na aplicação de técnicas inteligentes, a fim de se extrair os padrões de interesse do usuário. Para que essa etapa ocorra, é necessário que os dados estejam preparados através de etapas, tais como limpeza, seleção e transformação dos dados.

(B) as etapas de Pós-processamento e visualização dos Resultados ocorrem quando a técnica de Análise de Outliers é a escolhida para analisar os dados.

(C) Redes Neurais podem ser usadas na tarefa de classificação. Nesse caso, o algoritmo de classificação terá como entrada um banco de dados de treinamento e retornará como saída uma rede neural que contém uma árvore de decisão a ser analisada.

(D) na tarefa de análise de Agrupamentos (ou análise de clusters), são trabalhados dados que já foram classificados. Por isso, a tarefa consiste em identificar agrupamentos com regras de associação embutidas.

Finalizando Data Mining

Conceitos Complementares

Prof. Thiago Cavalcanti

Formas de mining

- **Preditivo** - A data mining pode mostrar como certos atributos dos dados irão se comportar no futuro
- **Textual -** Processo de obtenção de informação utilizando fontes de dados textuais. Aplicações em classificação automática de textos e busca de agrupamentos.
- **Espacial -** Processo de descoberta de padrões utilizando bancos de dados espaciais populados por mapas.

Conhecimento indutivo

- A data Mining apoia o **conhecimento indutivo**, que descobre **novas regras e padrões** nos dados fornecidos.
- O conhecimento pode ser representado de muitas formas:
  - Quando não estruturado, pode ser representado por **regras ou por lógica proposicional**.
  - Em uma forma estruturada, pode ser representado por **árvores de decisão, redes semânticas, redes neurais** ou hierarquias de classes ou frames.

Séries temporais

- Uma **série temporal** é uma coleção de observações feitas sequencialmente ao longo do tempo.
  - Em séries temporais a **ordem dos dados** é fundamental.
  - Uma característica muito importante deste tipo de dados é que **as observações vizinhas** são dependentes e o interesse é analisar e modelar esta dependência.

OLAP x Data Mining

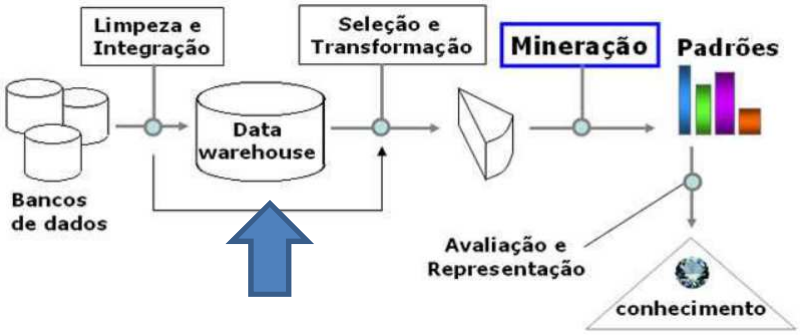
- OLAP
  - O termo para processamento analítico on-line representa
    - A característica de **trabalhar os dados** com **operadores dimensionais**
    - Possibilita uma **forma múltipla** e **combinada** de análise

OLAP x Data Mining

- Data Mining
  - Está mais relacionado com os **processo de análise de inferência** do que com a análise dimensional de dados
  - Representa uma forma de **busca de informação** baseada em algoritmos que objetivam o reconhecimento de padrões escondidos nos dados
    - **Não** necessariamente **revelados** pelas outras abordagens analíticas, como o **OLAP**

Data Mining x Data Warehousing

- DW – propõe sustentar a tomada de decisão com dados. (Armazém)
  - Uma coleção de dados **orientada por assunto**, **integrada**, **não-volátil**, **variante no tempo**, que dá apoio às decisões da administração.
- DM – pode ser usada em conjunto com o DW para auxiliar certos tipos de decisão



ITnerante

Mais uma definição pra finalizar

"A mineração de dados é um campo interdisciplinar que reúne técnicas de aprendizado de máquina, reconhecimento de padrões, estatísticas, banco de dados e visualização para abordar a questão da extração de informações a partir de grandes bases de dados"

(Evangelos Simoudis, citado em Daniel T. Larose, Discovering Knowledge in Data – An Introduction to Data Mining).

ITnerante

TIMASTERS

Data Mining

CESPE UnB
UNIVERSIDADE DE BRASÍLIA

FUNDAÇÃO CESGRANRIO

Fundação Carlos Chagas

ESAF
Escola de Administração Fazendária

FGV

Prof. Thiago Cavalcanti

Questão - Cargo 12: Técnico – Qualificação: Programação e Controle de Serviços de Tecnologia da Informação – Serpro 2013

Julgue os itens que se seguem à luz dos conceitos básicos de datamining e datawarehouse.

[110] Em tarefas preditivas, o atributo a ser predito é conhecido como variável independente, enquanto que os atributos usados para fazer a predição são conhecidos como alvo.

[111] Em algoritmos de clusterização hierárquica, os clusters são formados gradativamente por meio de aglomerações ou divisões de elementos, gerando uma hierarquia de clusters.

[112] Tarefas descritivas têm como objetivo derivar padrões como correlações, tendências, grupos, trajetórias e anomalias, os quais sumarizam as relações subjacentes nos dados.

[113] Nos métodos de particionamento para k-clusterização e k medoids, o elemento que melhor representa o cluster é definido de acordo com seus atributos, sem que haja muita influência dos valores próximos aos limites do cluster.

Questão - Cargo 13: Técnico – Qualificação: Programação e Controle de Serviços de Tecnologia da Informação – Serpro 2013

- Julgue os itens seguintes, relativos à arquitetura e às tecnologias de sistemas de informação.
- [89] Datamining é a tecnologia por intermédio da qual os processos são automatizados mediante racionalização e potencialização por meio de dois componentes: organização e tecnologia

Questão 14 - ESAF - 2012 - CGU - Analista de Finanças e Controle - prova 3 - Auditoria e Fiscalização - Geral

A Mineração de Dados é

(a) o processo de desenvolvimento de soluções automáticas de acesso a informações úteis em depósitos de dados.
(b) a transformação automática de dados existentes em grandes depósitos de dados em informações quantificáveis.
(c) a automação da recuperação de informações caracterizadas por registros com grande quantidade de atributos.
(d) a descoberta de relações significativas entre dados e informações passíveis de atualização automática.
(e) o processo de descoberta automática de informações úteis em grandes depósitos de dados.

Definição: Michael Berry, Gordon Linoff

- Data Mining Techniques: For Marketing, Sales, and Customer Support
  - 2011 - A mineração de dados é **um processo de negócio** para explorar grandes quantidades de **dados** para descobrir **padrões** e **regras significativas**.
  - 1997 - A mineração de dados é a **exploração e análise**, por meios **automáticos** ou **semiautomático**, de grandes quantidades de dados a fim de descobrir padrões e regras significativas.

Questão 15. ANA 2009 - Desenvolvimento de Sistemas e Administração de Banco de Dados

20- Para que uma empresa efetue a análise regular de dados gerados por visitantes do seu site *Web*, a fim de personalizar a propaganda para clientes individualmente, ela deve utilizar

a) objetos distribuídos.

b) mineração de dados.

c) processamento analítico online (olap).

d) diretório de informações.

e) sistema de informação gerencial.

Questão 16. ESAF – CVM 2010 - Sistemas

53- Mineração de Dados é

a) o processo de atualizar de maneira semiautomática grandes bancos de dados para encontrar versões úteis.

b) o processo de analisar de maneira semiautomática grandes bancos de dados para encontrar padrões úteis.

c) o processo de segmentar de maneira semiautomática bancos de dados qualitativos e corrigir padrões de especificação.

d) o programa que depura de maneira automática bancos de dados corporativos para mostrar padrões de análise.

e) o processo de automatizar a definição de bancos de dados de médio porte de maior utilidade para os usuários externos de rotinas de mineração.

Questão 17. BANESE - Técnico Bancário III - Área Informática – Desenvolvimento -2012

52. Data Mining é parte de um processo maior denominado

(a) Data Mart.

(b) Database Marketing .

(c) Knowledge Discovery in Database.

(d) Business Intelligence.

(e) Data Warehouse.

Questão 18. MPOG 2010 – Tecnologia da Informação

41- Mineração de Dados

a) é uma forma de busca sequencial de dados em arquivos.

b) é o processo de programação de todos os relacionamentos e algoritmos existentes nas bases de dados.

c) por ser feita com métodos compiladores, método das redes neurais e método dos algoritmos gerativos.

d) engloba as tarefas de mapeamento, inicialização e clusterização.

e) engloba as tarefas de classificação, regressão e clusterização.

Questão 19 – STN AFC 2008 - TECNOLOGIA DA INFORMAÇÃO/ INFRA-ESTRUTURA DE TI

13- Com respeito à mineração de dados, assinale a opção correta, após avaliar as seguintes afirmações:

I. A mineração de dados pode ser usada em conjunto com um *datawarehouse*, para auxiliar tomada de decisão.

II. A mineração de dados permite a descoberta de regras de associação entre hierarquias.

III. A mineração de dados compreende todo o processo de descoberta de conhecimento em bancos de dados.

a) Apenas as afirmações I e II são corretas.

b) Apenas as afirmações I e III são corretas.

c) Apenas as afirmações II e III são corretas.

d) As afirmações I, II e III são corretas.

e) As afirmações I, II e III são incorretas.

Questão 20. TRF - 4ª REGIÃO - Analista Judiciário - Tecnologia da Informação - 2010

Sobre data mining, é correto afirmar:
(a) É o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.
(b) Não requer interação com analistas humanos, pois os algoritmos utilizados conseguem determinar de forma completa e eficiente o valor dos padrões encontrados.

(c) Na mineração de dados, encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados", de forma a desconsiderar aquilo que é genérico e privilegiar aquilo que é específico.

(d) É um grande banco de dados voltado para dar suporte necessário nas decisões de usuários finais, geralmente gerentes e analistas de negócios.

(e) O processo de descobrimento realizado pelo data mining só pode ser utilizado a partir de um data warehouse, onde os dados já estão sem erros, sem duplicidade, são consistentes e habilitam descobertas abrangentes e precisas.

Questão 21. CESPE - 2013 - TRE-MS - Analista Judiciário - Análise de Sistemas

Prova(s): No que se refere a arquiteturas e aplicações de *data warehousing*, ETL, Olap e *data mining*, assinale a opção correta.

(a) As ferramentas Olap agregam recursos de armazenamento, gerenciamento e pesquisa de dados, os quais são primordialmente voltados para a tomada de decisões e BI (*business intelligence*).

(b) Um sistema ETL, que faz parte do processo de construção de um *data warehouse*, por ser voltado para a tomada de decisões, utiliza unicamente a DSL (*decision support language*), não suportando a SQL (*structured query language*).

(c) Em uma modelagem multidimensional do tipo *snow flake*, as métricas ficam inseridas nas dimensões.

(d) Em comparação com o ambiente transacional, o ambiente de *data warehouse*, devido à carga de dados com o ETL, deve estar mais voltado para inserção e atualização de dados do que para consultas.

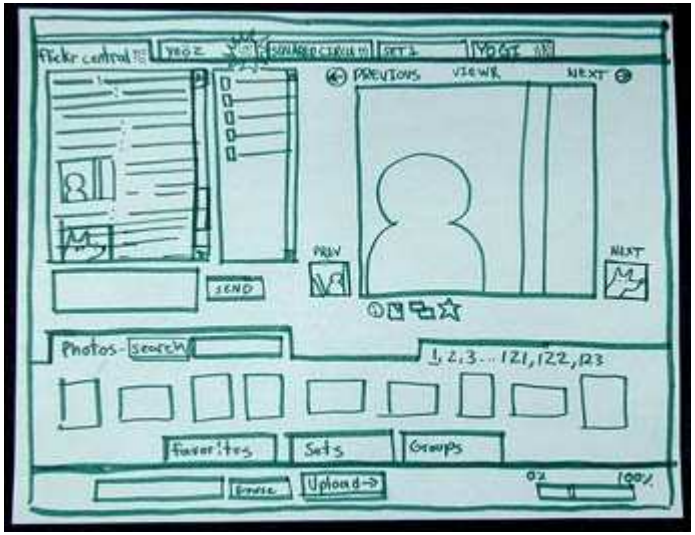
(e) *Data mining* é um conjunto de técnicas e ferramentas que permitem obter valores futuros a partir de dados passados processados estaticamente. *Data mining* substitui o *data warehouse* em relação à tomada de decisão, pois ambos possuem os mesmos recursos.

ITnerante

TIMASTERS

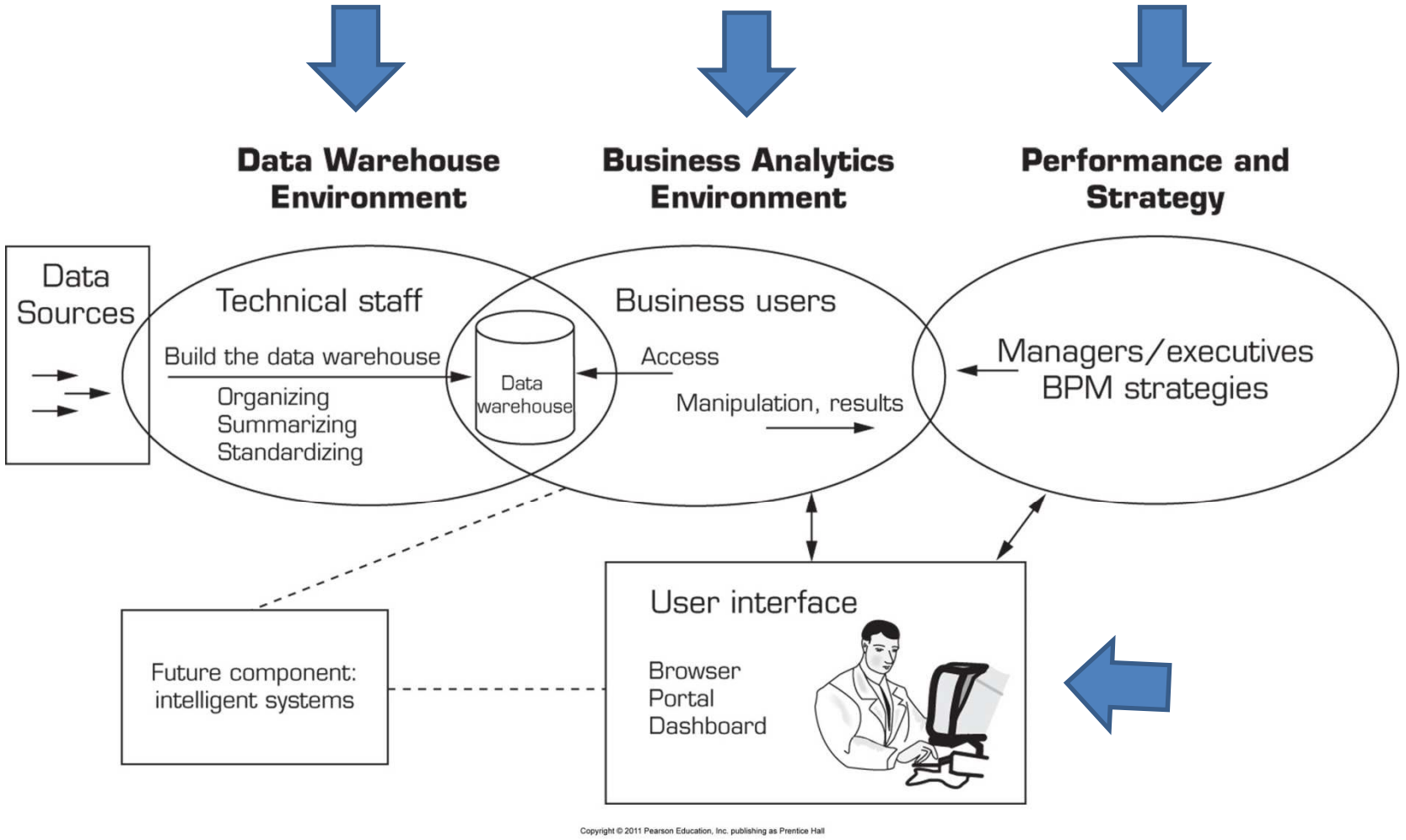
BPM e Interface com usuário

Business Performance Management



Prof. Thiago Cavalcanti

Componentes de Sistemas de BI



ITnerante

Conceito (Turban)

- Business Performance Management
  - Refere-se aos processos de negócios, metodologias, métricas e tecnologias utilizadas pelas empresas para medir, monitorar e gerenciar o desempenho dos negócios
  - É considerado o componente final do processo de Business Intelligence.
  - Utiliza **a análise**, **a geração de relatórios** e as **consultas de BI** com o objetivo de aperfeiçoar o desempenho geral da organização
  - Também conhecido como:
    - **Corporate performance Management**

Conceitos

- Interfaces de usuários
  - São as **ferramentas de visualização** que apresentam as informações de uma maneira compreensível aos usuários.
  - Estas podem ser **dashboards** (fornecem uma visão abrangente e amigável dos indicadores chaves de desempenho e suas tendências e exceções), **cubo multidimensional** de dados e, até mesmo, **realidade virtual**

BPM: Componentes

- Três componentes principais:
  1. Um **conjunto de processos** integrados, cíclico e analíticos, apoiado pela **tecnologia**, que trata de **atividades financeiras** bem como as **operacionais**
  2. Um **conjunto de ferramentas** para empresas **definirem objetivos estratégicos** e, em seguida, **medir e gerenciar** o desempenho
  3. Um **conjunto básico de processos**, incluindo planejamento financeiro e operacional, consolidação e relatórios, modelagem, análise e monitoramento de **indicadores-chave de desempenho (KPIs)**, vinculado à estratégia organizacional

Key Performance Indicator (KPI)

- É uma métrica associada a um processo.
- O KPI foca no **desempenho do processo**, o que representa a distância entre a medida e o alvo objetivado.
- Os KPI's têm por **objetivo** melhor comunicar a evolução dos indicadores por meio de uma pontuação e verificar se os resultados estão próximos ou longe do alvo.

Painéis e dashboards

Estrutura de BI

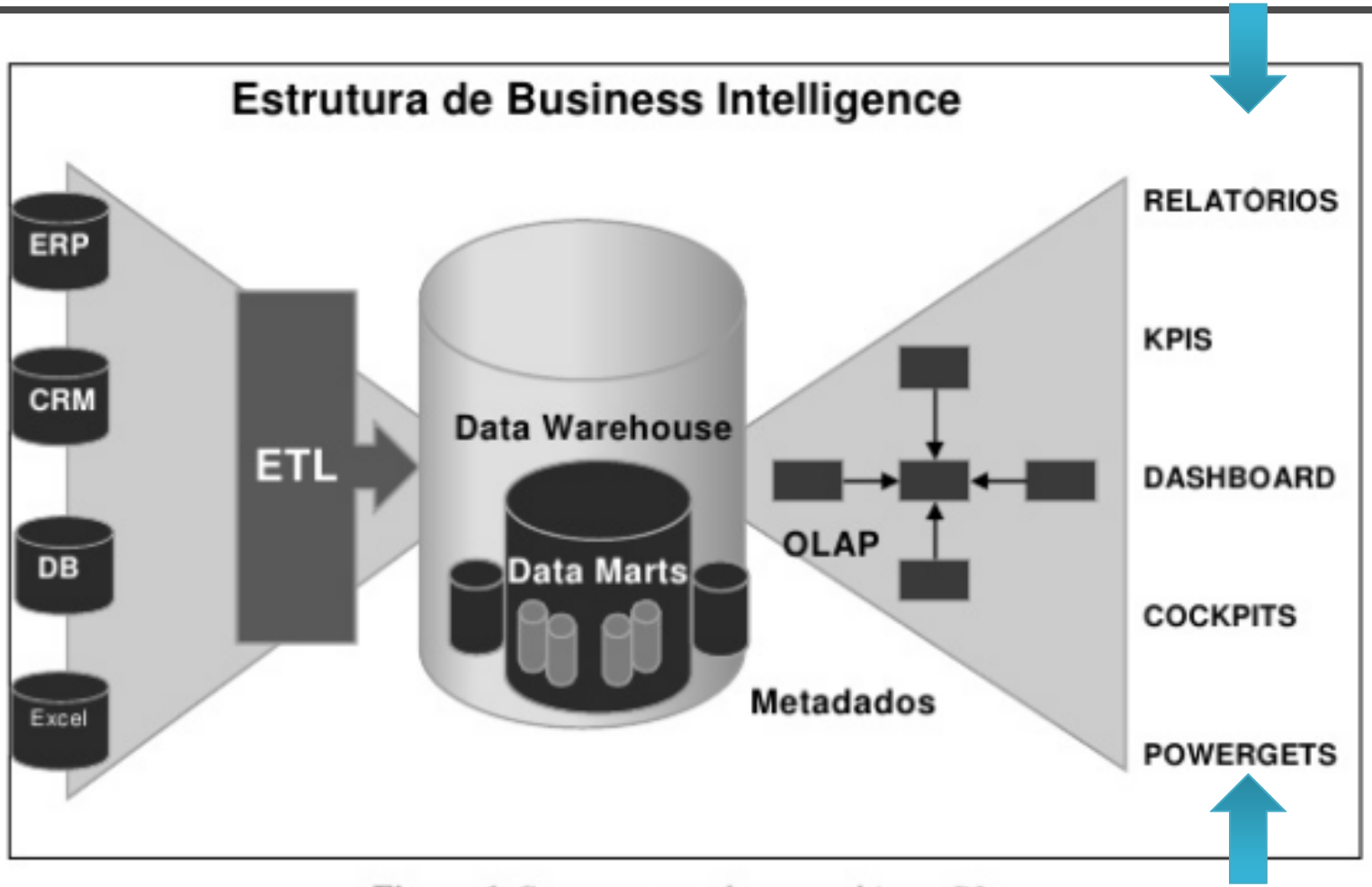


Figura 1 *Componentes de um ambiente BI.*
Fonte: Adaptação do autor segundo (BARBIERI, 2001)

ITnerante

Relatórios

- É a apresentação de dados transformados em informações formatadas e organizadas de acordo com requisitos de negócios específicos.
- São problemáticos quando demonstram apenas dados operacionais.
- Um relatório geralmente é estático, não permite que o visualizador tenha acesso a formatação dos dados, ao seu somatório ou a sua sumarização.
  - Em síntese, tendem a ser unidimensionais.
- Quando falamos em **BI**, esquecemos que existem relatórios, eles passam a ser vistos como **dashboards** e não mais como simples impressões de dados operacionais.

Dashboards e Scorecards

- Fornecem displays visuais de informações importantes que são consolidadas e dispostas em uma única tela
  - Desta forma a informação pode ser digerida em um único olhar e facilmente exploradas

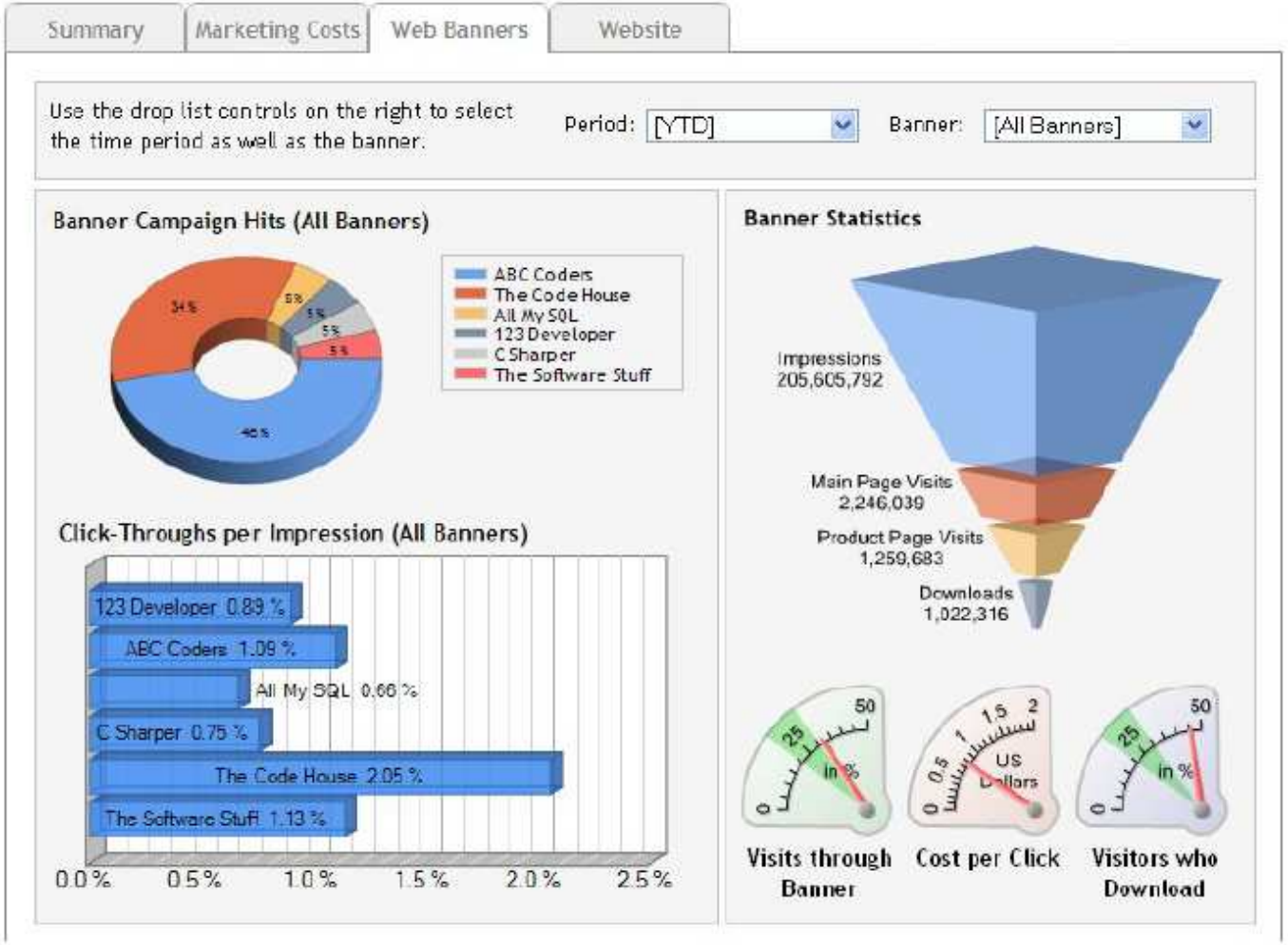
Dashboards x Scorecard

- Painéis
  - Exibição visual utilizada para **monitorar o desempenho** operacional (feito de **forma livre** ...)
- Scorecards
  - Exibição visual utilizada para traçar o progresso em relação às **metas e objetivos** estratégicos e táticos (medidas **predeterminadas** ...)

Dashboard

- Um **conjunto ou um grupo de visões analíticas** relacionado com tabelas de indicadores, relatórios, planilhas, gráficos e demais componentes de análise de informação.
- É uma coleção de vários itens podendo ser composto por várias **páginas ou abas**, contendo **análises diversas**.
- O essencial para se obter um excelente dashboard não está apenas na exposição dos dados, mas em fornecer ao usuário **um elevado nível de interação**.

Dashboard

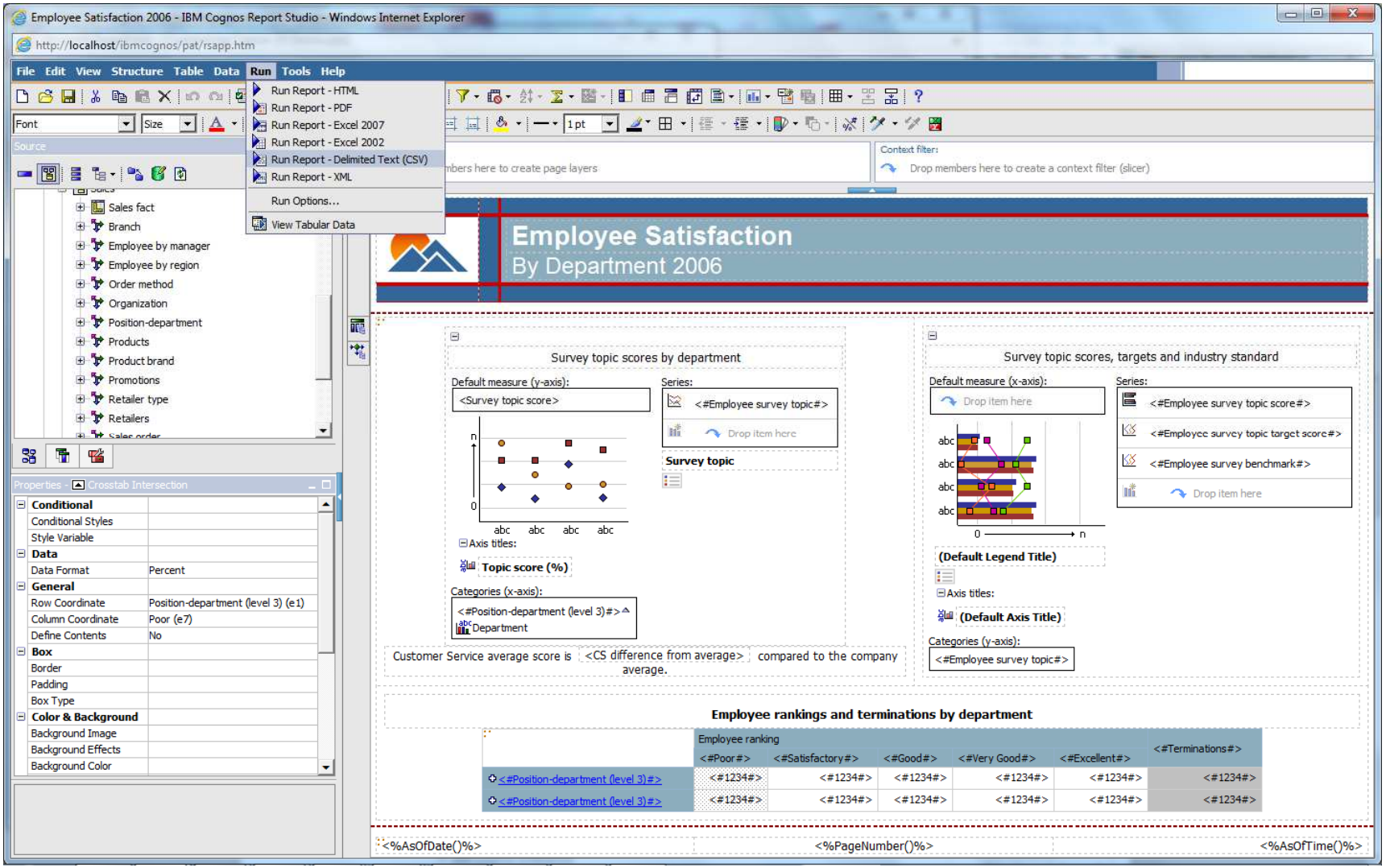


ITnerante

Cockpit

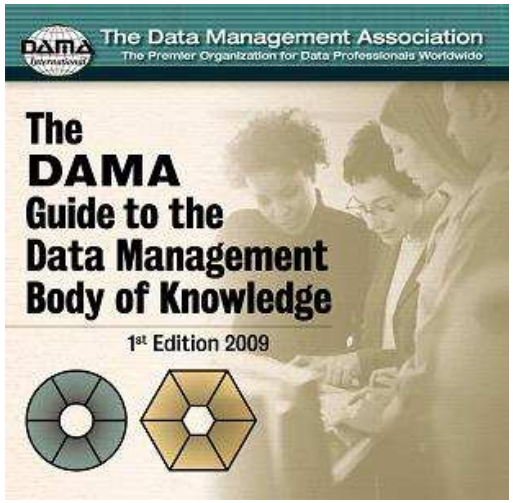
- O termo cockpit deriva da tradução de ‘cabina do piloto’.
- Um cockpit é formado geralmente por relógios, ou marcadores, que tem como função apresentar medidas de desempenho da empresa.
- No exemplo acima, podemos observar que os relógios indicam medidas sobre a eficiência da empresa

Powergets (IBM Congos PowerPlay)



ITnerante

Qualidade de dados

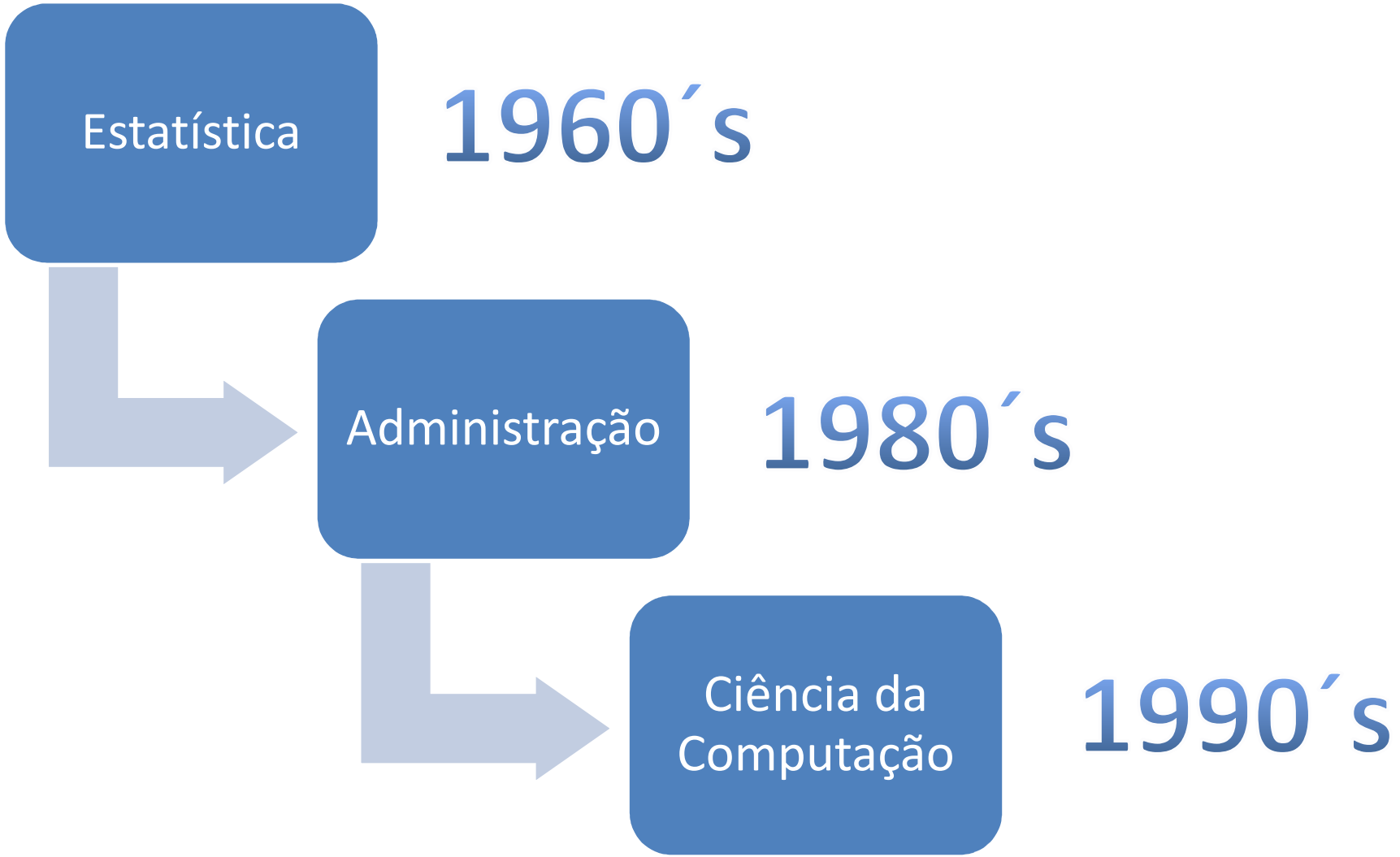


Uma introdução

Motivação

- A **qualidade dos dados** tem consequências graves, de relevância de longo prazo, para a eficiência e eficácia das organizações e empresas.
- O relatório sobre a qualidade do **Data Warehousing Institute** estima que:
  - os problemas de qualidade de dados custam empresas dos Estados Unidos mais de **600 bilhões de dólares** por ano

Histórico do conceito



Questionamentos

- Quais os problemas de qualidade de dados?
- Como detectar esses problemas?
- O que podemos fazer para resolver?

- Exemplos de problemas
  - Noise and outliers (ruídos e pontos for a da curva)
  - Valores faltantes
  - Dados duplicados.

Conceitos

- Noise (ruidos)
  - Refere-se a uma **modificação dos valores originais**
- Outliers
  - Objetos de dados com características que são **consideravelmente diferente** do que a **grande maioria** dos demais objetos no conjunto.

ITnerante

Valores Ausentes

- Razões para valores ausentes
  - Informação não coletada
    - Pessoas que não fornecem informações em uma pesquisa
  - Atributos que não podem ser aplicados a todos os casos
    - Renda anual não pode ser aplicado a crianças
- Gerenciando valores ausentes
  - Elimina-se objetos de dados
  - Estima-se os valores ausentes
  - Ignora-se valores durante a analise
  - Substitui-se por todos os valores possíveis (balanceando de acordo com as probabilidades)

Revisão da qualidade de dados

1. Identificar **os principais componentes** de dados associados com as políticas empresariais
2. Determinar como identificar o conteúdo dos dados pode **afetar o negócio**
3. Avaliar como **erros de dados são classificados** dentro de um conjunto de dimensões de qualidade de dados
4. Especificar as regras de negócio que **medem a ocorrência** de erros de dados
5. Fornecer um meio para **a implementação de processo de medição** que avalia a conformidade com essas regras de negócio.

Categorias dos dados

- Intrínseca: características intrínsecas dos dados, **independentes da sua aplicação**
- Acessibilidade: aspectos relativos ao **acesso e à segurança** dos dados.
- Contextual: características **dependentes do contexto** de utilização dos dados
- Representacional: características derivadas da **forma como a informação** é apresentada

Dimensões da qualidade dos dados

CATEGORIA	DIMENSÃO	DEFINIÇÃO
Intrínseca	Acuracidade (*acuracy* ou *free-of-error*)	Quanto a informação é correta e confiável
	Objetividade (*objectivity*)	Quanto a informação é imparcial
	Credibilidade *(believability)*	Quanto a informação é considerada como verdadeira e verossímil
	Reputação *(reputation)*	Quanto a informação considerada em termos de sua fonte ou conteúdo
Acessibilidade	Acessibilidade *(accessibility)*	Quanto a informação está disponível, ou fácil e rapidamente recuperável
	Segurança no acesso *(access security)*	Quanto o acesso a informação, é restrito apropriadamente para manter sua segurança
Contextual	Relevância *(relevancy)*	Quanto a informação é aplicável e útil para a tarefa a ser realizada
	Valor agregado *(value-added)*	Quanto a informação é benéfica e proporciona vantagens por seu uso
	Temporalidade/oportunidade *(timeliness)*	Quanto a informação está suficientemente atualizada para a tarefa a ser realizada
	Integridade/perfeição (completeness)	Quanto a informação não está extraviada e é suficiente para a tarefa em amplitude e profundidade
	Quantidade de informação apropriada *(appropriate amount)*	Quanto o volume da informação é apropriado para a tarefa ser executada

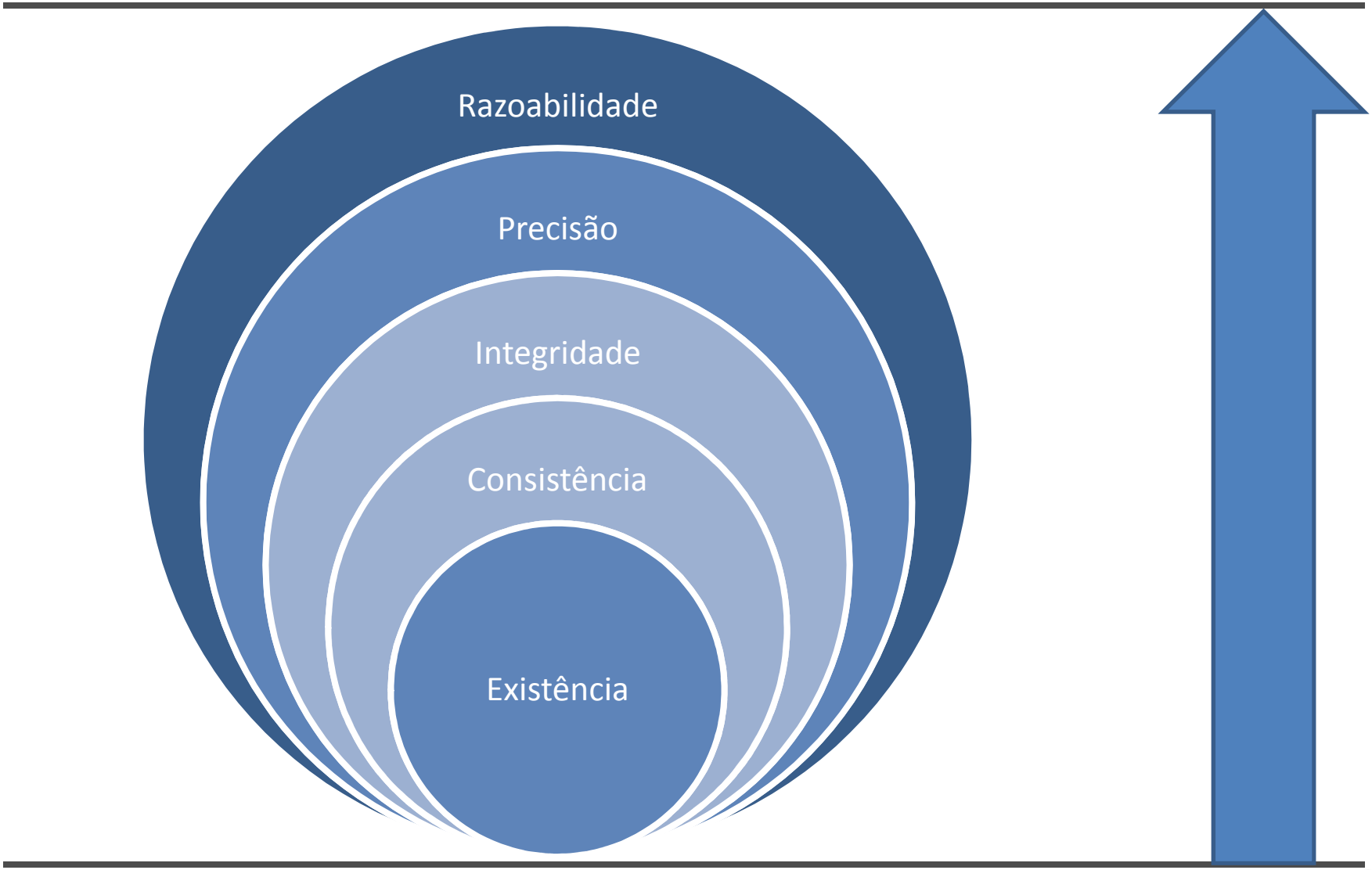
Dimensões da qualidade dos dados

CATEGORIA	DIMENSÃO	DEFINIÇÃO
Representação	Interpretabilidade *(interpretability)*	Quanto a informação está em linguagem apropriada, símbolos e unidades, e as definições são claras
	Facilidade de entendimento *(ease of understanding)*	Quanto a informação é facilmente compreendida
	Representação concisa *(concise representation)*	Quanto a informação está compactamente representada
	Representação consistente *(consistent representation)*	Quanto a informação é apresentada em um mesmo formato
	Facilidade de manipulação /operação *(ease of manipulation /operacion)*	Quanto a informação é fácil de ser manipulada e aplicada em diferentes tarefas

Qualidade de Dados

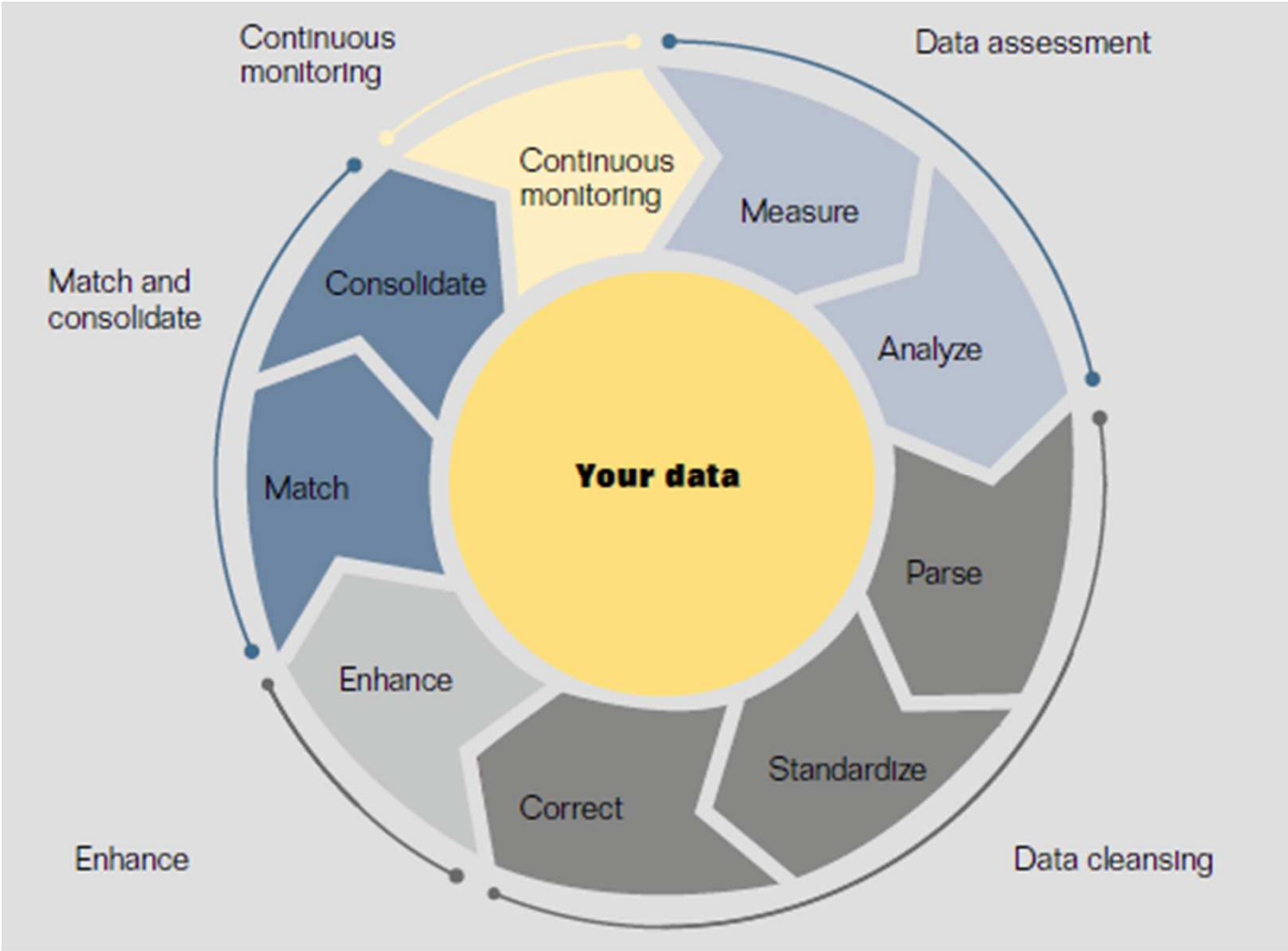
- A qualidade dos dados é um conceito multifacetado, cuja definição permite diferentes dimensões.
- As **dimensões de qualidade**, por exemplo, a **precisão**
  - Pode ser facilmente detectado **em alguns casos** (por exemplo, erros ortográficos), mas são mais difíceis em outros casos (por exemplo, onde são fornecidos os valores admissíveis, mas não correto).
- Tem sido mostrado um exemplo simples de um erro de **completude**, mas como com precisão, integridade também podem ser muito difícil de avaliar (por exemplo, se uma tupla representando um filme é totalmente ausente na relação Filme).
- **Detecção de consistência** nem sempre localiza os erros (por exemplo, para o filme 1, o valor ou o atributo LastRemakeYear é errado).

Uma visão temporal das dimensões



ITnerante

Processo simplificado de qualidade dos dados



ITnerante

Tipos de dados

- Estruturado - quando cada elemento de dados tem uma **estrutura fixa** associada.
  - Tabelas relacionais são o tipo mais popular de dados estruturados.
- Semiestruturado - Algumas características comuns são:
  - (i) os dados podem conter **campos não conhecidos** em tempo de design; por exemplo, um arquivo XML não tem um arquivo de esquema XML associado
  - (ii) **o mesmo tipo de dados pode ser representado em várias formas**; por exemplo, uma data pode ser representado por um campo ou por vários campos, mesmo dentro de um único conjunto de dados
  - (iii) entre os campos conhecidos em tempo de design, muitos campos não terão valores
- Não estruturado - quando os dados são **expressos em linguagem natural** e sem estrutura ou de domínio específico tipos são definidos

Dados agregados e elementares

- Dados elementares são geridos de organizações por processos operacionais e representam fenômenos atômicos do mundo real
  - Exemplo: número de segurança social, idade, sexo
- Os dados agregados são obtidos a partir de uma coleta de dados elementares através da aplicação de **uma função de agregação** para eles
  - Exemplo: a renda média dos contribuintes em uma determinada cidade
- Esta classificação é útil para distinguir diferentes níveis de gravidade para **medir e alcançar a qualidade dos dados**.
  - Como exemplo, **a precisão** de um atributo Sexo muda drasticamente se a entrada M (masculino) em vez de F (feminino);
  - se a idade de uma única pessoa é erroneamente registrado como 25 em vez de 35, a precisão da **idade média** de uma população de milhões de habitantes é **minimamente afetada**.

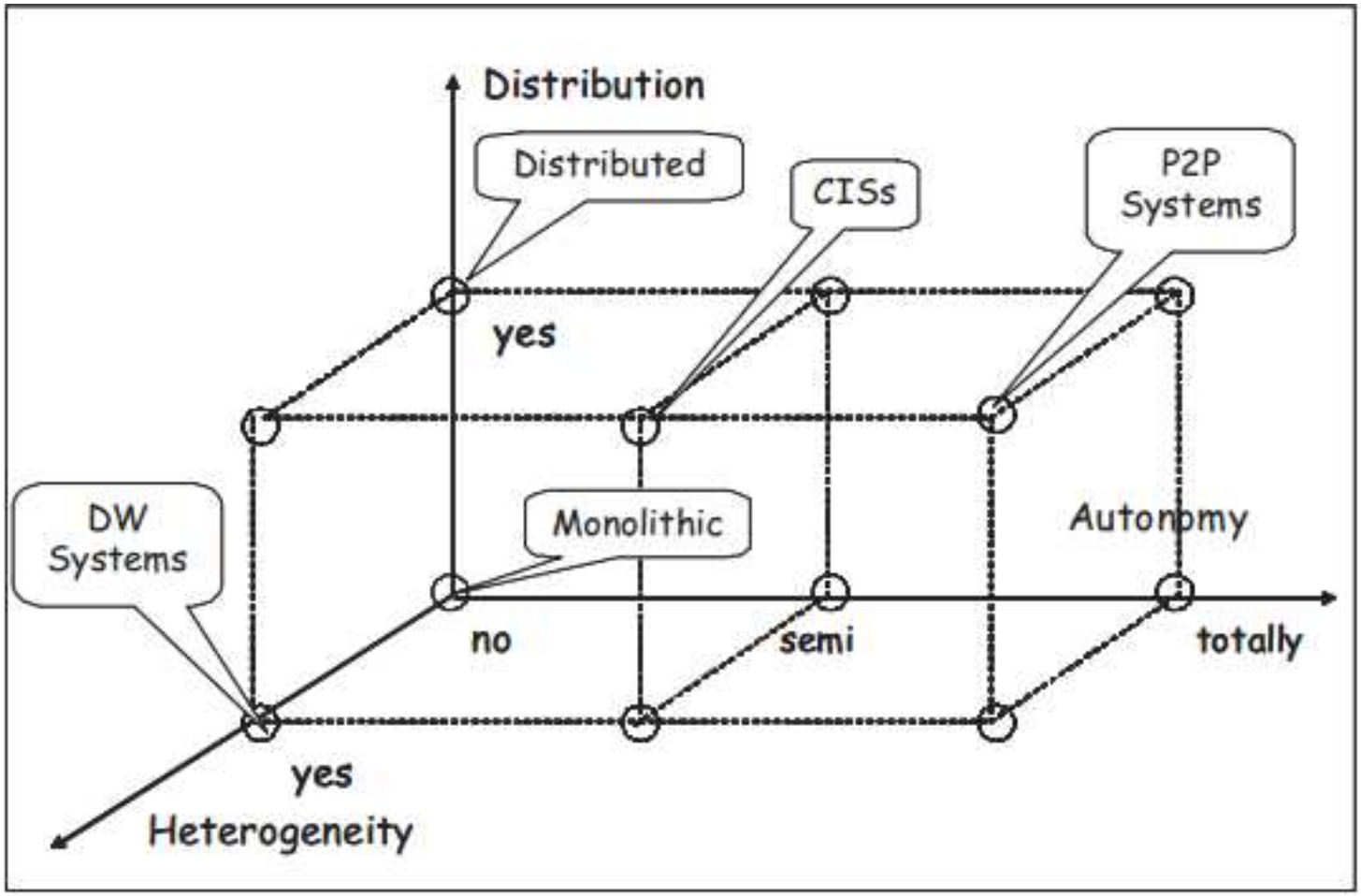
Outras classificações

- Dados estáveis x Dados que mudam no longo prazo x Dados que mudam frequentemente.
- A fim de esclarecer o impacto da qualidade dos dados sobre os **diferentes tipos de sistemas de informação**, que adaptam os critérios de classificação propostos para **bancos de dados distribuídos** são propostos três critérios diferentes:
  - (i) distribuição, (ii) heterogeneidade e (iii) autonomia

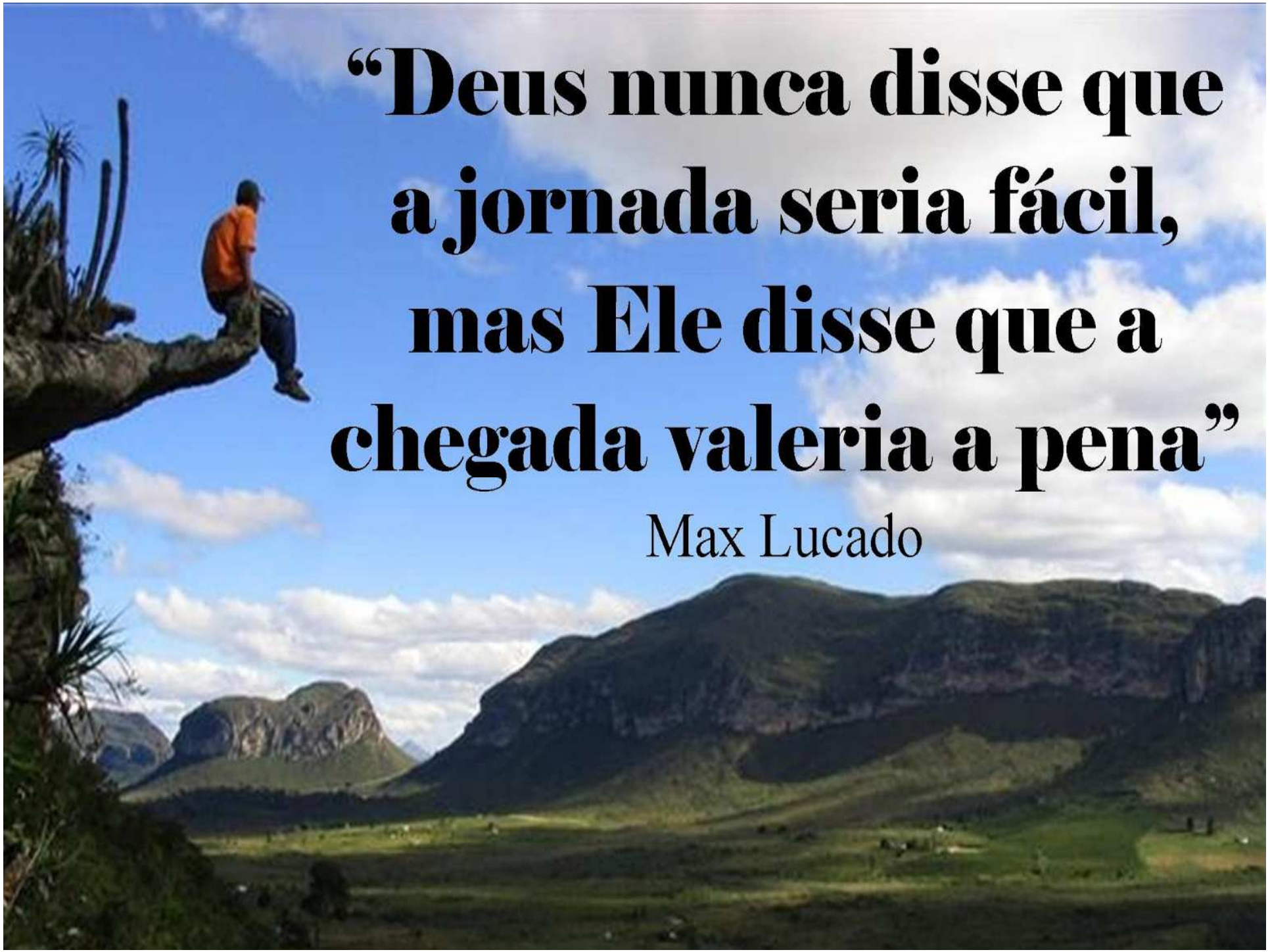
Distribuição, Heterogeneidade e Autonomia

- **Distribuição** – lida com possibilidade de **distribuir os dados e aplicação** em uma rede de computador
- **Heterogeneidade** - considera todos os tipos de **diversidades semânticas e tecnológicas** entre os **sistemas** utilizados na **modelagem e na representação** física dos dados, tais como o SGBDs, linguagens de programação, sistemas operacionais, middleware, linguagens de marcação.
- **Autonomia** - tem a ver com o **grau de hierarquia** e **as regras de coordenação**, estabelecendo direitos e deveres definidos na organização usando o sistema de informações.

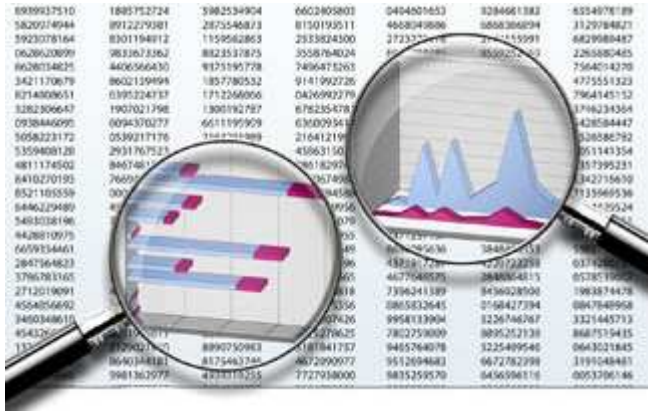
Tipos de sistemas da informação



ITnerante


“Deus nunca disse que a jornada seria fácil, mas Ele disse que a chegada valeria a pena”
Max Lucado

Valeu Galera!! :P



ITnerante

TIMASTERS

Banco de dados

Suporte a decisão – Business Intelligence

Curso Preparatório - ITnerante

Prof. Thiago Cavalcanti

Gabarito – Data Mining

1. C
2. D
3. A
4. C
5. A
6. B
7. C
8. D
9. C
10. A
11. A
12. [110] E [111] C [112] C [113] C
13. E
14. E
15. B
16. B
17. C
18. E
19. A
20. A
21. A