



SISTEMAS DE BANCO DE DADOS 2

AULA 14

Conhecendo melhor os Dados

Vandor Roberto Vilardi Rissoli



APRESENTAÇÃO

- Conhecendo mais os Dados
- Processar volume de dados
 - Tipos de dados
- Modelagem e representação
- Referências



CARACTERÍSTICAS DOS DADOS

A IMPORTÂNCIA DOS DADOS

- Dados sempre foram gerados desde tempos remotos a nossa realidade, porém o volume dessa geração nos tempos atuais é enorme.
- No início dessa década a **IBM** publicou que 90% de todos os dados do mundo foram gerados nos últimos 2 anos.



90% = últimos 2 anos

10% = desde o início



CARACTERÍSTICAS DOS DADOS

- Com tantos dados alguns desafios precisavam ser superados, destacando-se entre eles como:

- Armazenar tantos dados;
- Utilizar esses dados.



- O avanço de métodos e tecnologias coerentes a essa realidade têm sido fundamentais para tais superações

- *Data Warehouse*

- *Business Intelligence*

- *Big Data*

- *Data Mining*

- *Dashboard*

(visualização)



- entre outras.

CARACTERÍSTICAS DOS DADOS

HISTÓRIA DOS DADOS

- Dados são essenciais as pessoas e organizações;
- Guardados de diversas formas, mas evoluíram em seu armazenamento (físico => digital);
- Anteriormente, extrair informações e manter os dados organizados era uma tarefa muito **custosa**;
- O acesso aos dados dependia da **localização geográfica** dos arquivos que os armazenavam;



- Apenas armazenar não resolvia o problema, era necessário que os dados se relacionassem e pudessem ser usados pelas pessoas interessadas.



CARACTERÍSTICAS DOS DADOS



Fitas Magnética
Cartão Perfurado
Leitura de Dados
Sequencial



Edgar Frank Codd
propõe o modelo
de dados relacional.
E surge o termo
SGBDR.



Computador
Pessoal (PC)
Sistemas de Banco
de Dados
Linguagem SQL
Banco de Dados
Comerciais

1950 - 1960

1960 - 1970

1970 - 1980

1980

1980 - 1990



Discos Rígidos
Leitura Não
Sequencial
Modelo de Dados
Hierárquico
Modelo de Dados
em Rede

Dr. Peter Chen
propõe o modelo
Entidade-
Relacionamento



CARACTERÍSTICAS DOS DADOS

- Estima-se que diariamente são gerados 15 Petabytes de informações (redes sociais, dispositivos móveis, financeiros) em todo o mundo, provenientes de diversas plataformas e sistemas;



Facebook 10 e **Twitter** 7 Terabytes diários cada um;

Dados gerados nas **pesquisas astronômicas** armazenaram só em 2010 cerca de 140 Terabytes, tendo a expectativa desse volume de dados ser gerado a cada 5 dias com os novos **telescópios**.



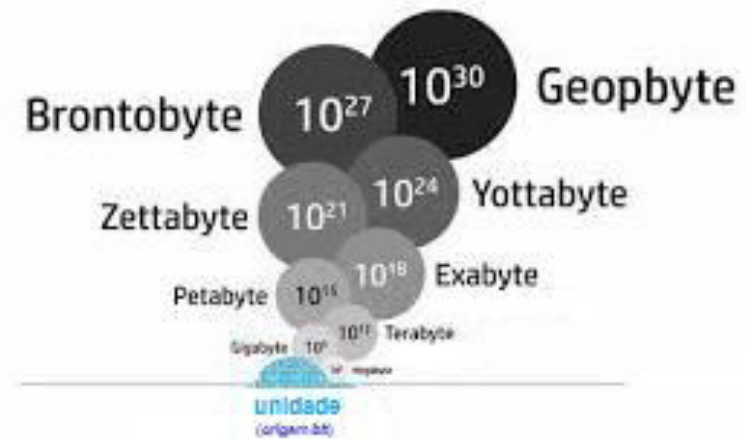
CARACTERÍSTICAS DOS DADOS

VOLUME DE DADOS

- No ano 2000, eram armazenados no mundo 800 mil Petabytes (PB)
- A expectativa da IBM para 2020 é a geração de 35 Zettabytes (ZB)

SISTEMA DE MEDIDAS EM ESPAÇO DE MÍDIAS ELETRÔNICAS

MEDIDA USUAL	Nº DE CARACTERES	BASE 2
Byte	1	2^0
Kilobyte (KB)	1.024	2^{10}
Megabyte (MB)	1.048.576	2^{20}
Gigabyte (GB)	1.073.741.824	2^{30}
Terabyte (TB)	1.099.511.627.776	2^{40}
Petabyte (PB)	1.125.899.906.842.624	2^{50}
Hexabyte (HG)	1.152.921.504.606.846.976	2^{60}



- Como processar esse volume enorme de dados?



<https://www.youtube.com/watch?v=hEFFCKxYbKM>

CARACTERÍSTICAS DOS DADOS

Observando as características dos dados armazenados nos anos indicados é possível notar que existem diferenças entre eles (**Tipos de Dados**).

1996 => 80% Textos simples, HTML
=> 20% Filmes, Figuras, Documentos

2012 => 80% Filmes, Figuras, Documentos
=> 20% Textos simples, HTML)



CARACTERÍSTICAS DOS DADOS

TIPOS DE DADOS

- Algumas características distinguem os tipos de dados existentes e possíveis de serem processados atualmente, sendo eles:

- **ESTRUTURADOS**

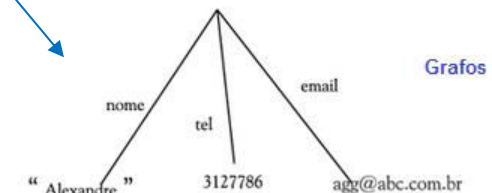
Nome	Idade	Departamento		Salário	Telefone
Alberto da Silva	25	Vendas	R\$	850,00	555-1902
Antônio dos Santos	32	Administração	R\$	1.200,00	555-1117
Fabiana Rossi	40	Administração	R\$	2.000,00	555-8929
Horácio Almeida	31	Recursos Humanos	R\$	1.350,00	555-8907
João Pereira	35	Vendas	R\$	1.500,00	555-7814
Roberto Albuquerque	29	Administração	R\$	1.200,00	555-8273
Sônia Pires	23	Vendas	R\$	600,00	555-8664

- **SEMIESTRUTURADOS**

```
{  
  pessoa :  
    {nome:"Alexandre",telefone:3127786,email:"agg@abc.com.br"},  
  pessoa:  
    {nome:"Sara",telefone:2136877,email:sar@math.com.br},  
  pessoa:  
    {nome:"Frederico",telefone:7734412,email:"fds@ac.co.kk"}  
}
```

Lista ou
Esquema

- **NÃO ESTRUTURADOS**
(ou desestruturados)



Grafos



CARACTERÍSTICAS DOS DADOS

DADOS ESTRUTURADOS

- Dados organizados em blocos semânticos para serem armazenados e manipulados (tabelas);
- Dados de um mesmo bloco possuem as mesmas descrições (atributos), sendo estas descrições agrupadas para formar estruturas (esquemas) que descrevem os seus blocos de dados relevantes;
- Dados mantidos em um SGBD são chamados de **Dados Estruturados** por manterem a mesma estrutura de representação (rígida), previamente projetada (esquemas ou tabelas).

Nome	Idade	Departamento		Salário	Telefone
Alberto da Silva	25	Vendas	R\$	850,00	555-1902
Antônio dos Santos	32	Administração	R\$	1.200,00	555-1117
Fabiana Rossi	40	Administração	R\$	2.000,00	555-8929
Horácio Almeida	31	Recursos Humanos	R\$	1.350,00	555-8907
João Pereira	35	Vendas	R\$	1.500,00	555-7814
Roberto Albuquerque	29	Administração	R\$	1.200,00	555-8273
Sônia Pires	23	Vendas	R\$	600,00	555-8664



CARACTERÍSTICAS DOS DADOS

DADOS SEMIESTRUTURADOS

- Dados que **NÃO** estão armazenados no SGBD, mas em *Data Lakes* (“guarda dados brutos”);
- Dados com organização bastante **HETEROGÊNEA** e pouca distinção entre Estrutura e Valor (dado);
- Essas características **DIFICULTAM AS CONSULTAS** sobre este tipo de dado **Semiestruturado**;
- Dados **Semiestruturados NÃO** são:
 - Estritamente “tipados” (~~estruturados~~);
 - Completamente desestruturados (~~não-estruturados~~);
 - Seriam um meio termo entre **Estruturados** e **Não Estruturados**.



CARACTERÍSTICAS DOS DADOS

- Dados Semiestruturados são indicados como sem esquema (*schemaless*) ou autodescritivos (*self-describing* – explicados por si próprios);
- Não existe uma separação entre os dados e o esquema, estando contidos como dados os próprios esquemas;
- Geralmente, há o “casamento” entre o SGBD e os Sistemas Documentais (manipulam documentos);

SGBD

Otimizações
B-tree, Hash

...



Sist. Documentos

Índices de texto
Classificação

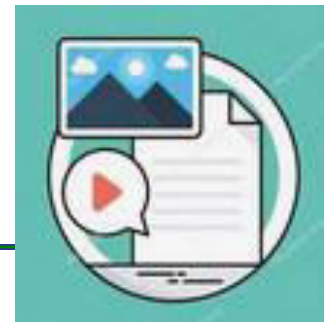
...



CARACTERÍSTICAS DOS DADOS

NÃO ESTRUTURADOS

- Dados que **NÃO** possuem uma estrutura definida para armazenamento (descrições **não** são comuns);
- Caracterizados, normalmente, por documentos textos, imagens, vídeos e outras “formas” de dados;
- Nem as estruturas são descritas implicitamente como nos Dados Semiestruturados;
- Grande maioria dos dados atuais na Web e nas organizações seguem este “formato” (estima-se que 80% de todos os dados armazenados no mundo são **Dados Não Estruturados**).



CARACTERÍSTICAS DOS DADOS

- Quando não é possível identificar uma organização clara dos dados armazenados, se reconhece estar diante de um dado **Não Estruturado**.

Por exemplo:

- Como identificar todas as palavras de um documento texto (bloco de notas, Word, e-mails, etc.) e relacioná-las em um contexto coerente?
- É praticamente impossível, não é? Quando nos deparamos com esta situação desorganizada estamos lidando com dados **Não Estruturados**.



CARACTERÍSTICAS DOS DADOS

- SGBD não contêm todas as informações possíveis sobre algo que esteja lá armazenado (representado);
- Organizar dados torna-os informação, exigindo que campos específicos (atributos) sejam preenchidos para que a manipulação deles possa ser automatizada;
- Mas documentos de vários tipos (texto, imagem, vídeo, etc.) correspondem a informações relevantes e não podem ser assim organizados/estruturados, por exemplo:



- Quando pessoas em redes sociais colocam suas emoções no que transmitem fica impossível captá-las de maneira real em dados estruturados (SGBD).

CARACTERÍSTICAS DOS DADOS

- Comparação entre as características dos tipos

Dados Estruturados	Dados Semiestruturados	Dados Não Estruturados
Esquema pré-definido	Nem sempre há um esquema	Não há esquema
Estrutura regular	Estrutura irregular	Estrutura irregular
Estrutura independente dos dados	Estrutura embutida nos dados	Pode não ter estrutura alguma
Estrutura reduzida	Estrutura extensa (particularidades de cada dado, visto que cada um pode ter uma organização própria)	Estrutura extensa (particularidades de cada dado, visto que cada um pode ter uma organização própria)
Fracamente evolutiva	Fortemente evolutiva (estrutura modifica-se com frequência)	Fortemente evolutiva (estrutura modifica-se com frequência)
Prescritiva (esquemas fechados e restrições de integridade)	Estrutura descritiva	Estrutura descritiva
Distinção entre estrutura e dados é clara	Distinção entre estrutura e dados não é clara	Distinção entre estrutura e dados não é clara



CARACTERÍSTICAS DOS DADOS

- Como então trabalhar com diferentes tipos de dados em volumes significativos, enormes ou não?



- Entender a diferença entre os diferentes tipos de dados pode significar o **sucesso do seu projeto** de gestão de dados, documentos e informações para a organização.

CARACTERÍSTICAS DOS DADOS

MODELAGEM

- Tornar o projeto viável depende de conhecer conceitos e a própria cultura do ambiente corporativo;
- Algumas soluções estão disponíveis (XML, Grafos, OEM, JSON, DTD, etc.), mas conhecer a realidade para preparar o ambiente adequado de tecnologias é relevante para solução proposta que contemple todas as necessidades e seus respectivos tipos de dados propícios.

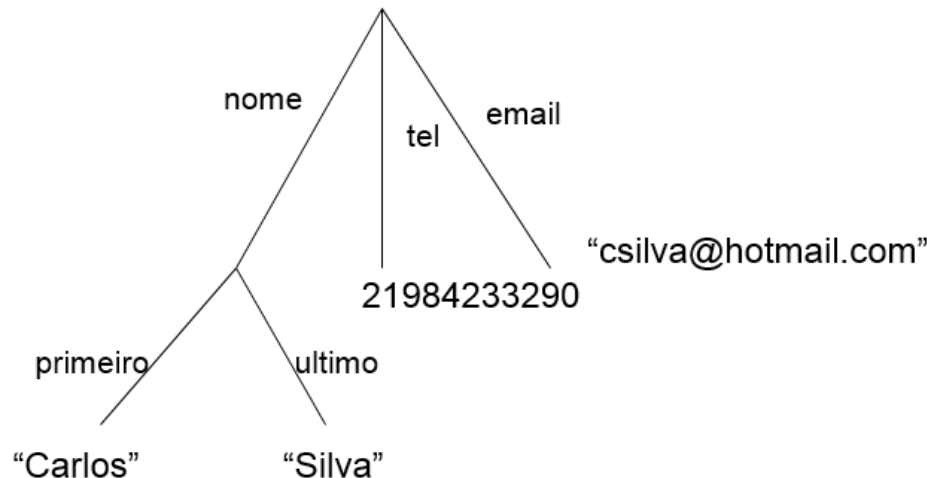


CARACTERÍSTICAS DOS DADOS

- Modelos que descrevem os dados de forma simples:

**{nome: {primeiro: "Carlos", ultimo: "Silva"},
tel: 21 984233290,
email: "csilva@hotmail.com"}**

- Representação em forma de Grafos:



CARACTERÍSTICAS DOS DADOS

- Os tipos de dados são esquecidos, deliberadamente, nos **Semiestruturados**, e são serializados os dados anotados com a suas descrições (autodescritivo):

{pessoa:

{nome: {primeiro: "Carlos", ultimo: "Silva"},
tel: 21 984233290,
email: "csilva@hotmail.com"},

pessoa:

{nome: "Paulo", tel: 21 34233267,
email: pavel@tagus.ist.utl.pt},

pessoa:

{nome: "Mara", tel: 11 44233246, peso: 62}

}

→ Acima está representado um objeto complexo.

CARACTERÍSTICAS DOS DADOS

- Um banco de dados relacional (Dados Estruturados) também pode ser descrito dessa forma simples:
- Suponha as tabelas **T1** e **T2** assim descritas por seus atributos **a**, **b**, **c**, **d**, **e**:

T1(a, b, c)

T2(d, e)

- No esquema de uma tabela (relação), cada atributo é definido sobre um domínio de valores atômicos;

{ **T1**: { linha: { **a**:a1, **b**:b1, **c**:c1 },
 linha: { **a**:a2, **b**:b2, **c**:c2 } } },

T2: { linha: { **c**:c2, **d**:d2 },
 linha: { **c**:c3, **d**:d3 } } }

T1		
a	b	c
a1	b1	c1
a2	b2	c2

T2	
c	d
c2	d2
c3	d3



CARACTERÍSTICAS DOS DADOS

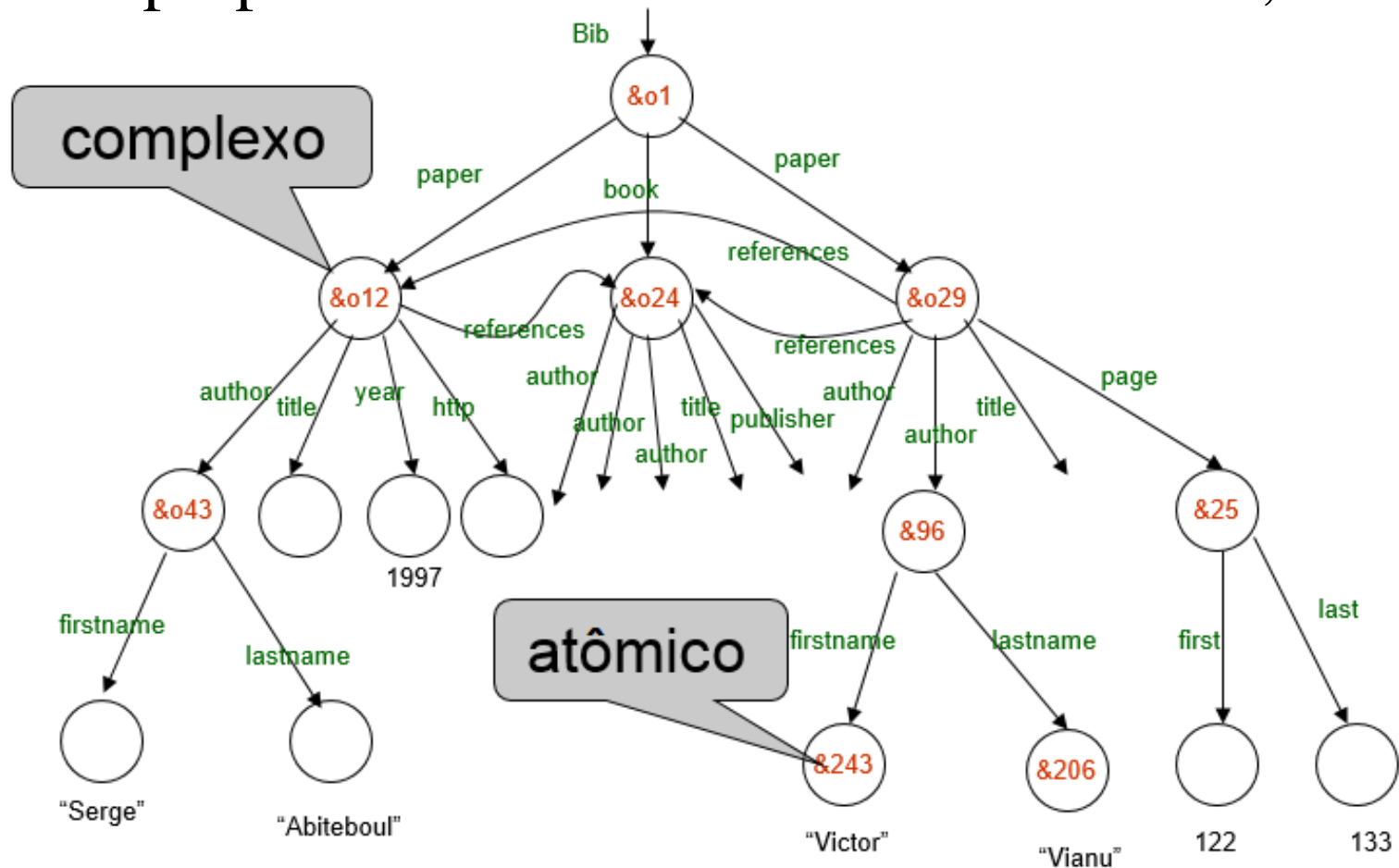
MODELO DE INTERCAMBIO DE OBJETO (OEM)

- OEM (*Object Exchange Model*) propicia a integração de fontes de dados heterogêneas;
- Um objeto OEM consiste em uma quadrupla (*label, oid, type, value*):
 - Etiqueta: cadeia de caracteres;
 - Identificado do objeto (IOD);
 - Tipo Complexo ou Atômico (int, string, gif, jpeg,...);
 - Valor (dado): *se* tipo Complexo, conjunto de OIDs;
senão um valor Atômico.



CARACTERÍSTICAS DOS DADOS

- OEM é um modelo de grafo;
- Os etiquetas estão nos nodos, mas várias extensões foram proposta e são aceitas também nos arcos;



CARACTERÍSTICAS DOS DADOS

XML - eXtensible Markup Language

- Uma linguagem de descrição de documentos, definida por um organismo internacional W3C;
- Um conjunto de tecnologias derivadas:
 - Xlink, Xpointer, Xschema, DOM, SAX, XSL,...
- O esperanto da Web;
- Origem no SGML (*Standard Generalized Markup Language*)
 - HTML (*HyperText Markup Language*) – descreve a apresentação
 - XML – descreve o conteúdo



CARACTERÍSTICAS DOS DADOS

- Exemplo de código HMTL simples;

<h1> Bibliography </h1>

<p> <i> Foundations of Databases </i>

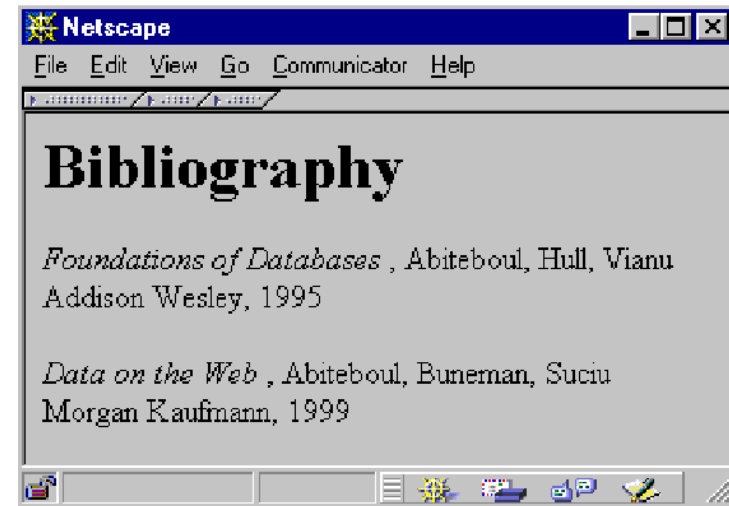
Abiteboul, Hull, Vianu

**
 Addison Wesley, 1995**

<p> <i> Data on the Web </i>

Abiteoul, Buneman, Suciu

**
 Morgan Kaufmann, 1999**



CARACTERÍSTICAS DOS DADOS

- Exemplo de código XML simples que descreve um conteúdo;

```
<bibliography>
```

```
  <book>  <title> Foundations... </title>
```

```
    <author> Abiteboul </author>
```

```
    <author> Hull </author>
```

```
    <author> Vianu </author>
```

```
    <publisher> Addison Wesley </publisher>
```

```
    <year> 1995 </year>
```

```
  </book>
```

```
  ...
```

```
</bibliography>
```



CARACTERÍSTICAS DOS DADOS

Exemplo

Como exemplo suponha que um sítio virtual de e-commerce precisa emitir uma Nota Fiscal, sendo o desafio atender os diversos departamentos da organização que usam estes dados das Notas Fiscais em diferentes plataformas, Sistemas Operacionais e linguagens de programação de diferentes aplicações.

- Para descrever o conteúdo das Notas Fiscais em diferentes ambientes seriam interessante usar XML.



CARACTERÍSTICAS DOS DADOS

<?xml version="1.0">

<NotaFiscal>

<NomeCliente> Ana Silva </NomeCliente>

<EndCliente> Rua Sul, Maceió, AL

</EndCliente>

<EndEnvio> Rua Sul, Maceió, AL </EndEnvio>

<Item>

<codigo> 123 </codigo>

<descricao> Parafuso 8mm </descricao>

<quantidade> 20 </quantidade>

<preco> 3,00 </preco>

</Item>

<Item> ... </Item>

</NotaFiscal>

CARACTERÍSTICAS DOS DADOS

<ficha>

<nome>

<fn>Vera</fn>

<ln>Santos</ln>

</nome>

<trab tipo="gerente">

IBGE

<end>

<cidade>Natal</cidade>

<cep>52310</cep>

</end>

<email>vera@ibge.br</email>

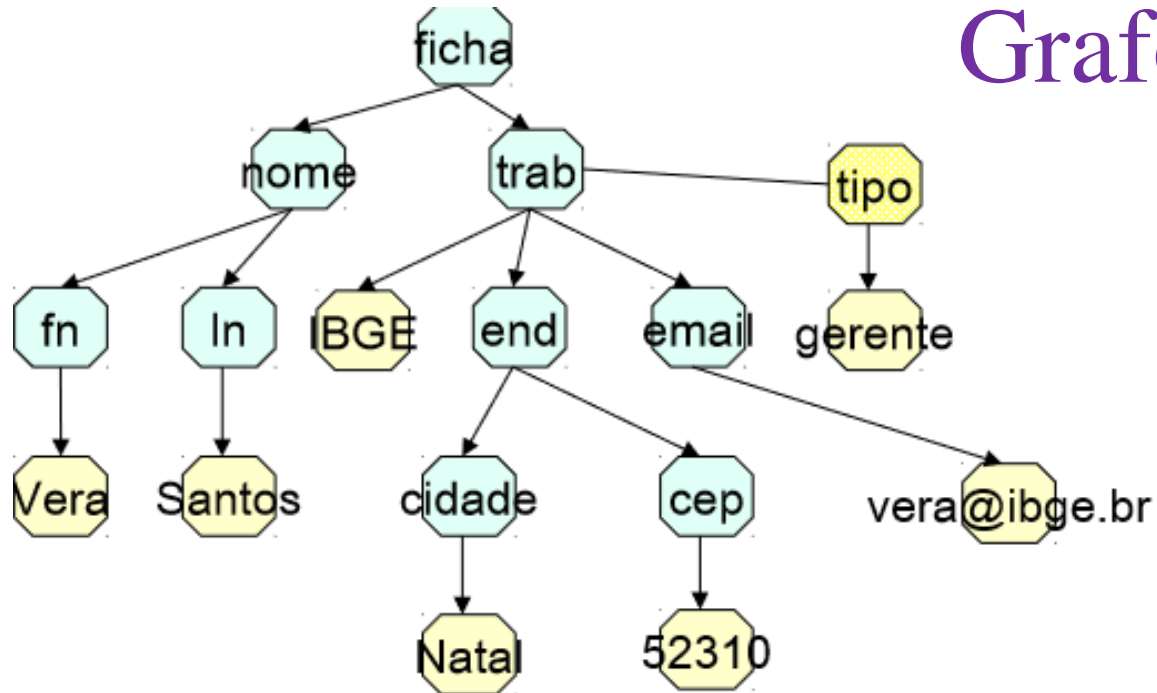
</trab>



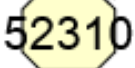
</ficha>

Sintaxe

CARACTERÍSTICAS DOS DADOS

Grafo



Elementos 
Atributos 
Dados 



CARACTERÍSTICAS DOS DADOS

VANTAGENS COM XML

- **Extensibilidade e Estrutura** (cria TAGs livremente conforme a necessidade e elabora estruturas que melhor facilitem as ações e intercâmbio de dados);
- **Interoperabilidade** - todos dados são vistos como documentos XML e não mais como arquivos em diferentes formatos;
- **Modularidade e Reutilização** (cada usuário é livre para definir suas próprias estruturas de documento e possível automatização com uso de padrões, por exemplo DTD);
- **Acesso a fontes de dados Heterogêneas** (formato de troca normalizado e independente de plataforma – simples editor de texto pode tratar dados da organização);

CARACTERÍSTICAS DOS DADOS

- A realidade demanda de tal necessidade de armazenamento contínuo, com dados que representem a realidade de maneira fidedigna aos acontecimentos, desejos e com agilidade;
- A manipulação desse universo de dados armazenados também necessita de agilidade para atender em tempo hábil as demandas existentes.
- O que seria a *Fast Data* diante dessa realidade?



<https://www.youtube.com/watch?v=ltNsfcmk1bg>



Referência de Criação e Apoio ao Estudo

Material para Consulta e Apoio ao Conteúdo

- SILBERSCHATZ, A. e KORTH, H. F. Sistemas de Banco de Dados, tradução da 6ª edição
 - Capítulo 20
- WIKILIVROS – SQL/ Dados Estruturados, Semiestruturados e Não Estruturados
 - Site:
https://pt.wikibooks.org/wiki/SQL/Dados_Estruturados,_Semi-Estruturados_e_N%C3%A3o_Estruturados
- Universidade de Brasília (UnB Gama)
 - Site: <https://cae.ucb.br/conteudo/unbfga>
(escolha a disciplina **Laboratório de Banco de Dados**)

