Desenvolvimento de Software para Persistência

Tipos de Dados:

Estruturados, semiestruturados e não estruturados

Prof. Regis Pires Magalhães
regismagalhaes@ufc.br

UFC

Dados estruturados, semiestruturados e não estruturados

- Dados estruturados
  - Representados em um formato estrito.
  - Ex: registro em uma tabela relacional segue o mesmo formato dos outros registros da tabela.
  - Para dados estruturados, projeta-se cuidadosamente o esquema de banco de dados a fim de definir a estrutura do banco de dados.
  - O SGBD verifica se todos os dados seguem as estruturas e restrições especificadas no esquema.

Dados estruturados, semiestruturados e não estruturados

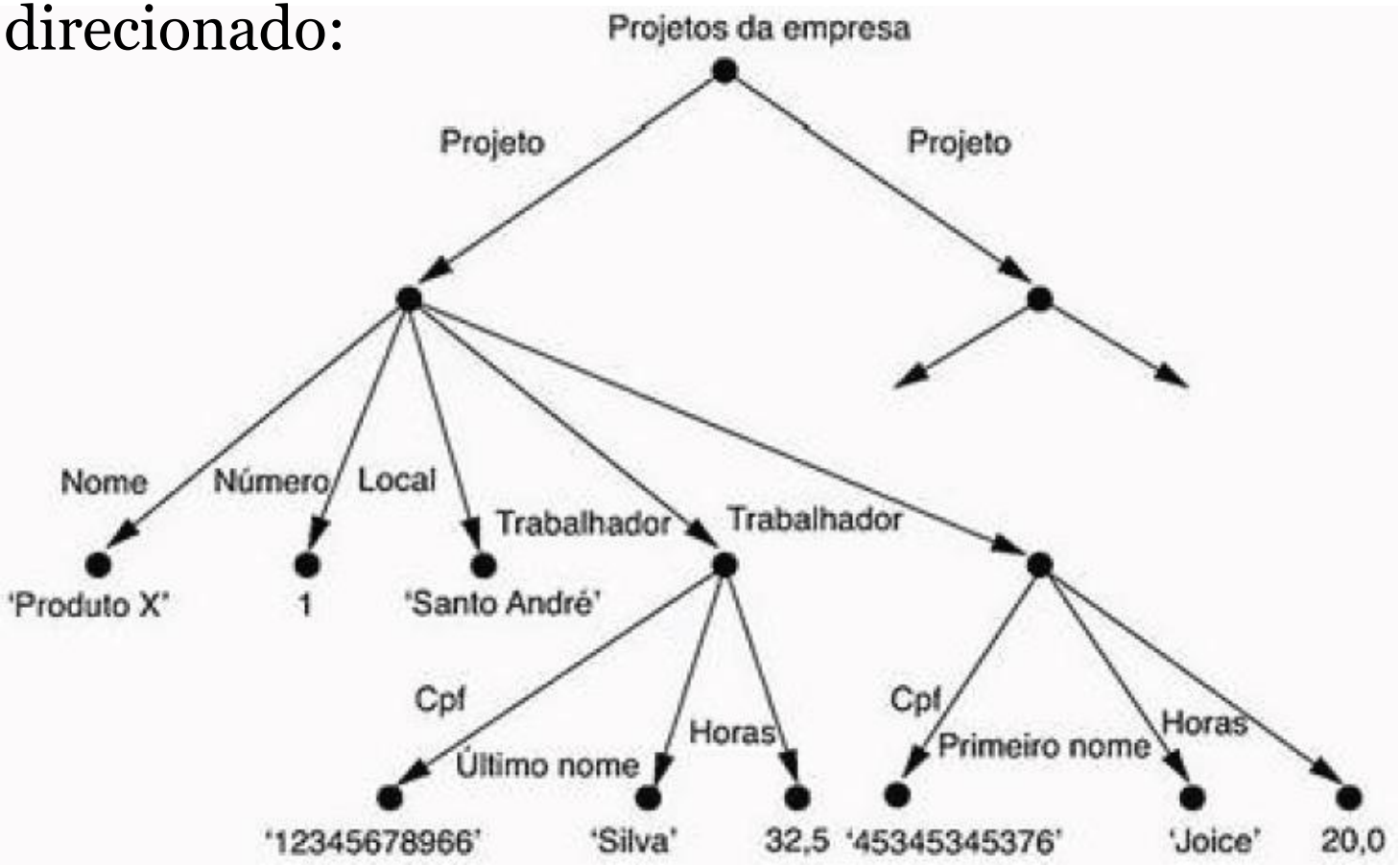
- Dados semiestruturados
  - Os dados podem ter uma estrutura, mas nem toda a informação coletada terá a estrutura idêntica.
  - Alguns atributos podem ser compartilhados entre as diversas entidades, mas outros podem existir apenas em algumas entidades.
  - Atributos adicionais podem ser introduzidos em alguns dos itens de dados mais novos a qualquer momento, e não existe esquema predefinido.
  - Diversos modelos de dados foram introduzidos para representar dados semiestruturados, geralmente com base no uso de estruturas de dados de árvore ou grafo, em vez das estruturas do modelo relacional plano.

Dados estruturados, semiestruturados e não estruturados

- Dados semiestruturados
  - Nos dados semiestruturados, a informação do esquema é misturada com os valores dos dados, já que cada objeto de dados pode ter atributos diferentes que não são conhecidos antecipadamente.
  - Logo, esse tipo de dado às vezes é chamado de dados autodescritivos.
  - Os dados semiestruturados podem ser exibidos como um grafo direcionado.
  - Esse modelo é capaz de representar objetos complexos e estruturas aninhadas.

Dados estruturados, semiestruturados e não estruturados

- Dados semiestruturados representados como um grafo direcionado:



Dados estruturados, semiestruturados e não estruturados

- Dados semiestruturados
  - A informação do esquema (nomes de atributos, relacionamentos e classes) no modelo semiestruturado é misturado com os objetos e seus valores de dados na mesma estrutura de dados.
  - No modelo semiestruturado não existe requisito para um esquema predefinido ao qual os objetos de dados precisam se adequar, embora seja possível definir um esquema, se necessário.

Dados estruturados, semiestruturados e não estruturados

- Dados não estruturados
  - Além de dados estruturados e semiestruturados, existe uma terceira categoria, conhecida como dados não estruturados porque existe indicação muito limitada sobre o tipo de dados.
  - Ex: documento de texto com informações incorporadas a ele.
    - As páginas HTML que contêm alguns dados são consideradas dados não estruturados.
    - As tags HTML especificam informações, como parágrafos, níveis de cabeçalho nos documentos, etc.
    - Algumas tags oferecem estruturação de texto, como na especificação de lista numerada ou não numerada, ou de tabela.
    - Mesmo essas tags de estruturação especificam que os dados textuais devem ser exibidos de certa maneira, em vez de indicar o tipo de dado representado na tabela.

Scraping Tools

- jsoup: Java HTML Parser
  - https://jsoup.org/
  - API para extração e manipulação de dados de documentos HTML.

```
Document doc = Jsoup.connect("http://example.com/").get();
String title = doc.title();
```

```
File input = new File("/tmp/input.html");
Document doc = Jsoup.parse(input, "UTF-8",
"http://example.com/");

Elements links = doc.select("a[href]"); // a with href
Elements pngs = doc.select("img[src$=.png]");

Element masthead = doc.select("div.masthead").first();

Elements resultLinks = doc.select("h3.r > a");
```

Extração a partir de PDF

- Tabula
  - http://tabula.technology
  - https://github.com/tabulapdf/tabula

Referências

- Elsmari, R., Navathe, Shamkant B. “Sistemas de Banco de Dados”. 6ª Edição, Pearson Brasil, 2011.

Obrigado!

Dúvidas, comentários, sugestões?

Regis Pires Magalhães

regismagalhaes@ufc.br