Integração de Dados

Departamento de Engenharia Informática e de Sistemas
Licenciatura em Engenharia Informática
2º ano/2º semestre

1

1

Resumo

- O que é a Integração de Dados
- Dificuldades da Integração de Dados
- Arquiteturas de Integração de Dados
- Componentes de um Sistema Integração de Dados
- Tarefas de um Sistema de Integração de Dados

2

2

Porque integrar?

3

3

Integração de Dados - para que serve?

- Permitir a extração, tratamento dos dados de múltiplas fontes de dados
  - **Heterogéneas**
  - **Distribuídas**
  - **Autónomas**
- Eliminar/corrigir redundâncias, conflitos, inconsistências

4

4

Integração de Dados

- Para que serve a **integração de dados**?
  - permitir pesquisas numa **vista unificada** e obter respostas válidas e consistentes
    - detectar correspondência entre conceitos similares
    - eliminar inconsistências
    - resolver conflitos
    - eliminar redundâncias

5

5

Integração de Dados: dificuldades

- Dificuldades da **integração de dados**?
  - **ao nível dos sistemas**
  - **ao nível da lógica**
  - ao nível administrativo e social
  - ao nível das expectativas

6

6

Integração de Dados: dificuldades

- Dificuldades da **integração de dados**?
  - **ao nível dos sistemas**
    - dados armazenados em sistemas com diferente hardware, diferente sistema operativo
    - dados armazenados em sistemas que usam linguagens de pesquisa diferentes
      - SQL, XPath, XQuery, ...
    - dados **distribuídos**: problemas no acesso
      - *firewalls*, protocolos de acesso/comunicação, autenticação, ...

7

7

Integração de Dados: dificuldades

- **ao nível da lógica**
  - Organização dos dados
  - Tipos de dados e valores
  - Semântica

8

8

Integração de Dados: dificuldades

- **ao nível da lógica**
  - Organização dos dados
    - Modelo relacional: esquemas (que podem ser diferentes!)
    - Outros modelos: tags XML, classes, propriedades
    - Dados sem estrutura

9

9

Integração de Dados: dificuldades

- **ao nível da lógica**
  - Tipos de dados e valores
    - dados incompletos e/ou inconsistentes
    - os mesmos dados, em fontes diferentes, podem ser representados de forma distinta:
      - tipo (e.g. datas representadas como String ou Date)
      - valor: F/M ou Feminino/Masculino; 8:00pm ou 20:00:00
      - valores numéricos representando moedas diferentes: 150 (EUR ou USD?)

10

10

Integração de Dados: dificuldades

- **ao nível da lógica**
  - Semântica
    - Os mesmos atributos podem ter diferentes significados em origens distintas (e.g. **titulo -- livro?, filme?, CD?, ...**)
    - Os mesmos dados podem estar em atributos com nomes diferentes (**nomeCli**, **nome**)

11

11

Integração de Dados: dificuldades

- Dificuldades da **integração de dados**?
  - ao nível administrativo e social
    - autorizações, burocracia, dados não digitais, documentação ....
  - ao nível das expectativas
    - compromisso entre a situação ideal e o que é possível fazer

12

12

Integração de Dados: dificuldades

A Web revolucionou a forma como os dados são gerados e manipulados!

- **Enorme quantidade de fontes de dados.**
- **Dados muito dinâmicos**
- **Dados bastante heterogéneos.**
- **Dados podem ser não estruturados ou semiestruturados**

13

13

Integração de Dados: uma boa solução

- Uma boa solução de integração de dados deve:
  - Permitir:
    - uma interface uniforme para as diversas fontes de dados
    - transparência de localização, modelo de dados e linguagem de consulta
  - Evitar:
    - a interação com cada origem de dados de forma isolada
    - o tratamento manual dos dados vindos dessas múltiplas fontes

14

14

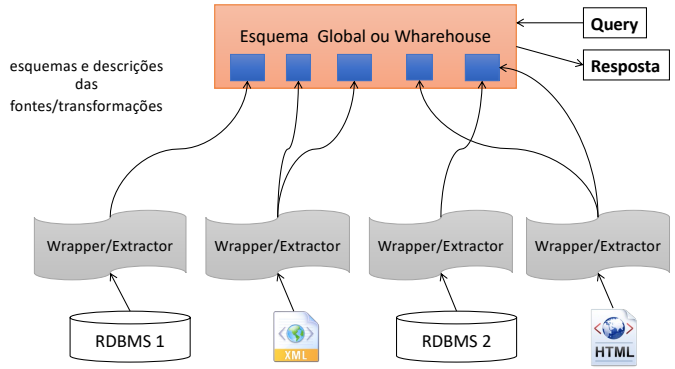
Integração de Dados: Arquitecturas

- Data warehousing (materializada)
  - Extrai os dados para uma fonte de dados comuns
- **Centralizada (virtual)**
  - Os dados permanecem nas suas fontes
    - abordagem que vamos aprofundar!

15

15

Integração de Dados: Arquiteturas

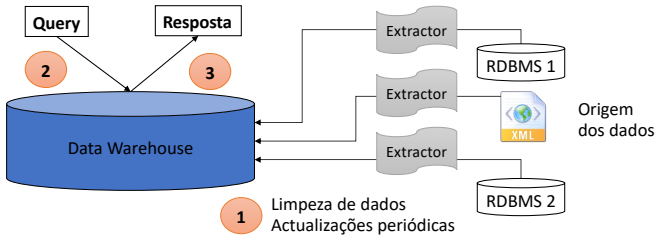


16

16

Integração de Dados: Arquiteturas

- **Data Warehouse:** Extrai os dados para uma fonte de dados comuns



17

17

Integração de Dados: Arquiteturas

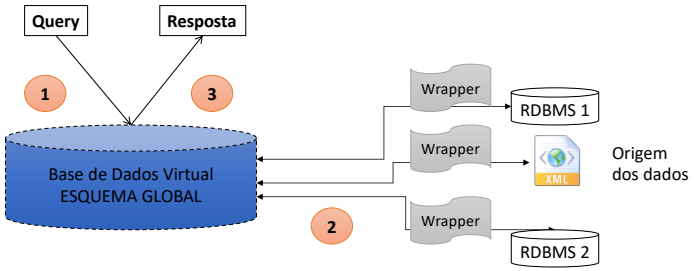
- **Data Warehouse**
  - Requere **limpeza** dos dados: diferentes formatos
    - ETL (Extract, Transform, Load)
  - Requere armazenamento dos dados em duplicado (BDs de origem e DW)
  - Requere atualizações periódicas dos dados:
    - As fontes de dados são autónomas – o conteúdo pode ser alterado sem aviso
  - Processo custoso a nível de limpeza de dados e espaço de armazenamento

18

18

Integração de Dados: Arquiteturas

- **Centralizada:** usa um esquema global



19

19

Integração de Dados: Arquiteturas

- Integração Centralizada (virtual)
  - Os dados permanecem nas suas fontes
  - Quando uma pesquisa/query é executada:
    - Determinam-se as fontes relevantes para dar resposta à *query*
    - Divide-se a *query* em diferentes *sub-queries* para as fontes dados
    - Obtêm-se as diferentes respostas e combinam-se de forma apropriada
  - Os dados estão sempre atualizados

20

20

Integração de Dados: componentes

- Um Sistema de Integração de dados I consiste em:

  I = <G, S, M>

  G = Esquema Global

  S = Esquemas das fontes de dados

  M = Mapeamentos entre S e G

21

21

Integração de Dados: componentes (arquitectura virtual)

- S (Fontes de dados):
  - onde reside a informação (heterogénea, autónoma): bases de dados relacionais, ficheiros html, xml, txt, aplicações
  - cada fonte de dados possui o seu esquema
- G (Esquema global)
  - esquema usado pelo utilizador para obter dados das várias fontes
  - possui apenas os aspectos relevantes para a aplicação
  - não armazena informação, apenas a descrição lógica
- M (Mapeamento semântico)
  - como os atributos de S se relacionam com os atributos de G
  - como resolver diferenças de valores em fontes distintas
  - concretizados com ***Wrappers***: enviam queries para as fontes de dados, processam as respostas, fazem transformações

22

22

Integração de Dados: tarefas - GLOBAL

I = <G, S, M>

- Analisar os modelos de dados de **S**
- Decidir como organizar os dados em **G**
- Definir os mapeamentos **M**
  - evitar inconsistências
  - evitar conflitos
  - eliminar redundâncias
  - permitir a execução de pesquisas

23

23

Integração de Dados: tarefas – 1. Analisar S

- **Resolver inconsistências/conflitos**
  - ao nível dos esquemas (atributos, tabelas, ...)
    - Identificar atributos idênticos:
      - `empregado, funcionario`
      - `disciplina, unidadeCurricular`
      - `cod, num`
    - Atributos **homónimos**
      - `Produtos(preco, ...)  -> Preço sem IVA`
      - `Produtos(preco, ...)  -> Preço com IVA`
    - Tabelas/Ficheiros **equivalentes**
      - `Escola(...)  ==  Estabelecimento(...)`
      - `Cliente(...) == Pessoas(...)`

24

24

Integração de Dados: tarefas – 1. Analisar S

- **Resolver conflitos**
  - ao nível dos valores
    - género: **`Masculino/Feminino – M/F – 0/1`**
    - identificação de pessoas**: `11233445333`** corresponde a BI, CC, NIF, outro?
  - ao nível da semântica
    - **`preço = 10`** **--- USD, EUR, … ?**
    - **`peso = 30`** **--- Kg, libras, … ?**
    - **`temperatura = 25`** **--- centigrados, farhenheit?**

25

25

Integração de Dados: tarefas – 1. Analisar S

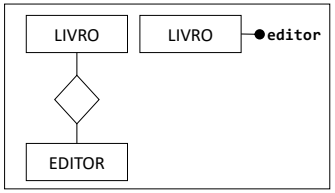
- **Resolver conflitos**
  - ao nível das chaves
    - `Produto(`**`cod`**`, ...)`
    - `Produto(`**`linha, cod`**`, ...)`

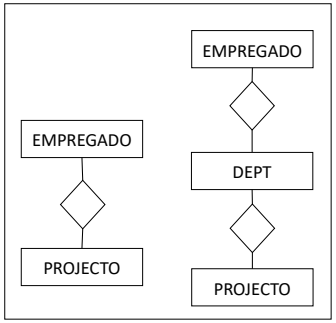
26

26

Integração de Dados: tarefas – 1. Analisar S

- **Resolver conflitos**
  - ao nível da estrutura dos esquemas



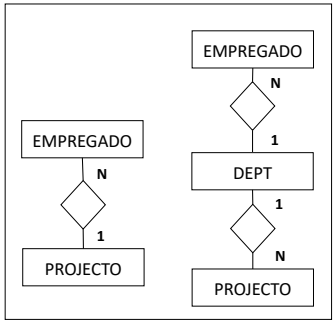


27

27

Integração de Dados: tarefas – 1. Analisar S

- **Resolver conflitos**
  - ao nível da dependência (cardinalidade) dos esquemas



28

28

Integração de Dados: tarefas – 2. Definir G

- **Definir o Esquema Global**
  - que tabelas/atributos interessam colocar em G
  - como organizar os dados
    - quantos esquemas?
    - como se relacionam?
    - que tipo de validação?

29

29

Integração de Dados: tarefas – 3. Definir M

- **Definir os Mapeamentos**
  - Para cada esquema de G, onde estão os dados relevantes em S
    - que tabelas ou ficheiros?
  - Para cada atributo de G quais os atributos homónimos de S
  - Que projecções de S devem ser criadas em G

30

30

Integração de Dados: pesquisas

- Efetuar pesquisas (queries)
  - **Reformulação** da *query* para poder aceder aos dados de acordo com S: **Q = Q1 + Q2 + ...**
  - **Optimização** da *query*:
    - qual a ordem pela qual as fontes S devem ser acedidas,
    - como são combinados os resultados
  - **Execução** da *query*:
    - Execução do plano: utilização dos **Wrappers** em S e junção dos dados obtidos

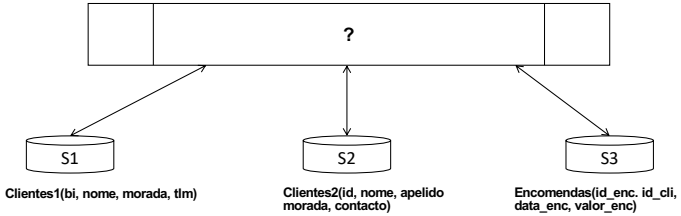
31

31

Integração de Dados: exemplo

Objetivo da pesquisa:

- **introduzir o id de um cliente**
  - **obter nome/contactos**
  - **obter lista de encomendas (datas, totais)**



32

32

Integração de Dados: exemplo

- **S**: 3 fontes de dados
  - **S1: armazena informação sobre clientes**
    - bi = identificação
    - nome = nome do cliente
    - morada = morada do cliente
    - tlm = contacto do cliente
  - **S2: armazena informação sobre clientes**
    - id = identificação (bi? Outra?)
    - nome + apelido = nome do cliente
    - morada = morada do cliente
    - contacto= contacto do cliente
  - **S3: armazena informação sobre encomendas feitas pelos clientes**
    - id_enc = identificação da encomenda
    - id_cli = identificação do cliente que fez a encomenda
    - data_enc = quando foi realizada a encomenda
    - valor_enc = valor total da encomenda

33

33

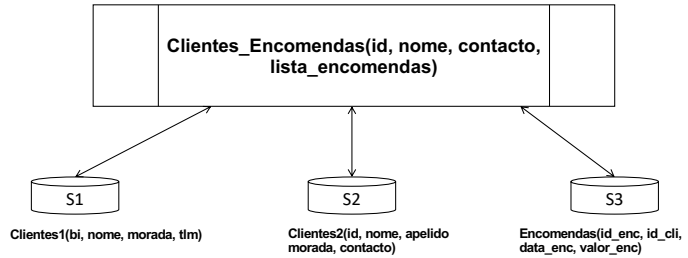
Integração de Dados: exemplo

- **G**: 1 tabela *(haveria outras possibilidades?)*
  - Clientes_Encomendas
    - id: *conflitos, redundâncias*
    - nome: *conflitos, redundâncias*
    - contacto: *conflitos, redundâncias*
    - lista_encomendas

34

34

Integração de Dados: exemplo



35

35

Integração de Dados: exemplo

- **M**: Mapeamentos S – G
  - **Clientes_Encomendas** construída a partir de S1, S2, S3
    - **id:**
      - bi de S1
      - id de S2
      - id_cli de S3
    - **nome:**
      - nome de S1
      - nome, apelido de S2
    - **morada:**
      - morada de S1 e de S2
    - **contacto:**
      - tlm de S1
      - contacto de S2
    - **lista_encomendas:**
      - id_enc, data_enc, valor_enc de S3

  >> extracção dos dados usando id_cli

36

36

Integração de Dados: Wrappers

- Como extrair dados das fontes?
  - Implementar Extractores/Wrappers
    - Usar linguagens de pesquisa específicas (SQL, Xquery, ...)
    - Usar Expressões regulares

37

37

Integração de Dados: Wrappers

- Tipos de *Wrappers*:
  - **manuais**: construídos após análise das fontes de dados e dos mapeamentos; são específicos para o esquema nesse dado momento
  - **automáticos**: usam técnicas de aprendizagem automática para se adaptarem, corrigirem e verificarem se continuam funcionais

38

38

Integração de Dados: Wrappers

- Limitações dos *Wrappers*:
  - **Manuais**: se o esquema da fonte de dados mudar, o código do *wrapper* tem de ser actualizado
  - Automáticos: complexidade, tempo de execução

39

39

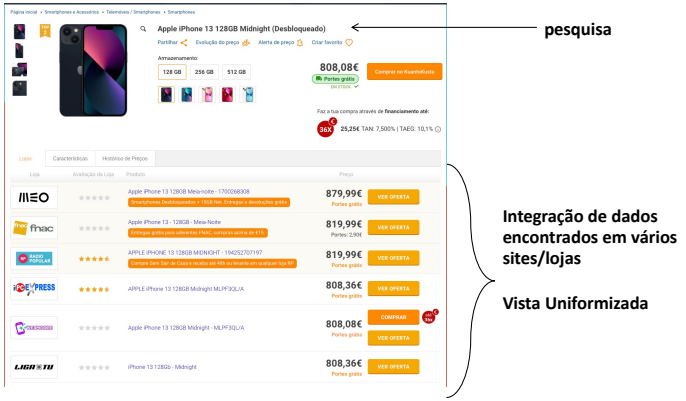
Integração de Dados: Wrappers

- Como implementar um *Wrapper*:
  - Pesquisas em SQL
  - **Expressões regulares**
    - construir expressões que encontrem padrões
      - ex: encontrar datas, emails, nºs de identificação, …
  - **Tecnologias XML**
    - uniformizar os diferentes formatos
    - visualizar os dados da forma desejada:
      - html
      - txt
      - xml
      - …

40

40

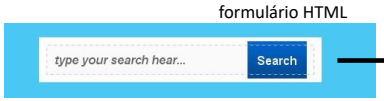
Integração de Dados: Wrappers



41

41

Integração de Dados: Wrappers – exemplo 1



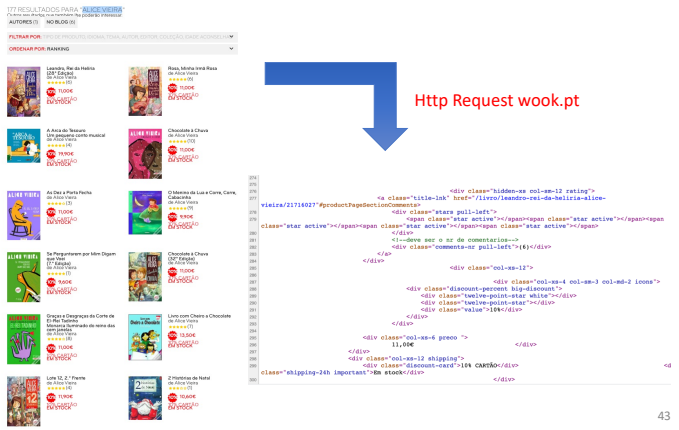
WRAPPER

- transforma a *query* em HTTP *requests*
- pesquisa fontes (HTML)
- transforma a resposta (dados semi estruturados) em tuplos
- devolve a informação para o formulário HTML

42

42

Integração de Dados: Wrappers – exemplo 1



43

43

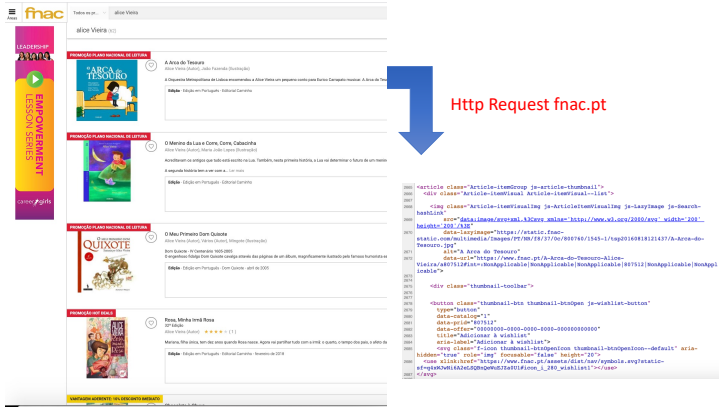
Integração de Dados: Wrappers – exemplo 1

Extrair dados WOOK:
<Titulo, preço, ISBN, Idioma, ...>

44

44

Integração de Dados: Wrappers – exemplo 1



45

45

Integração de Dados: Wrappers – exemplo 1

```
<article class="Article-itemGroup js-article-thumbnail">
  <div class="Article-itemVisual Article-itemVisual--list">

    <img class="Article-itemVisualImg js-ArticleItemVisualImg js-LazyImage js-Search-
hashLink"
        src="data:image/svg+xml,%3Csvg xmlns='http://www.w3.org/2000/svg' width='200'
height='200'/%3E"
        data-lazyimage="https://static.fnac-
static.com/multimedia/Images/PT/NR/E8/37/0c/800760/1545-1/tsp20160818121437/A-Arca-do-
Tesouro.jpg"
        alt="A Arca do Tesouro"
        data-url="https://www.fnac.pt/A-Arca-do-Tesouro-Alice-
Vieira/a807512#int=:NonApplicable|NonApplicable|NonApplicable|807512|NonApplicable|NonAppl
icable">

    <div class="thumbnail-toolbar">

    <button class="thumbnail-btn thumbnail-btnOpen js-wishlist-button"
      type="button"
      data-catalog="1"
      data-prid="807512"
      data-offer="00000000-0000-0000-0000-000000000000"
      title="Adicionar à wishlist"
      aria-label="Adicionar à wishlist">
      <svg class="f-icon thumbnail-btnOpenIcon thumbnail-btnOpenIcon--default" aria-
hidden="true" role="img" focusable="false" height="20">
    <use xlink:href="https://www.fnac.pt/assets/dist/nav/symbols.svg?static-
sf=q4xWJvWi6A2eLGQBnQeWuEJEaOUi#icon_i_280_wishlist1"></use>
</svg>
```

Wrapper 2

Extrair dados FNAC:
<Titulo, preço, ISBN, Idioma, ...>

46

46

Integração de Dados: Wrappers – exemplo 1

- Wrapper 1 // Wrapper 2
  - Ambos encontraram a informação solicitada?
  - Formatos ISBN são consistentes?
  - Há redundâncias / conflitos?

47

47

Integração de Dados: Wrappers - exemplo 1a

Objectivo: Inserir ISBN e obter preço

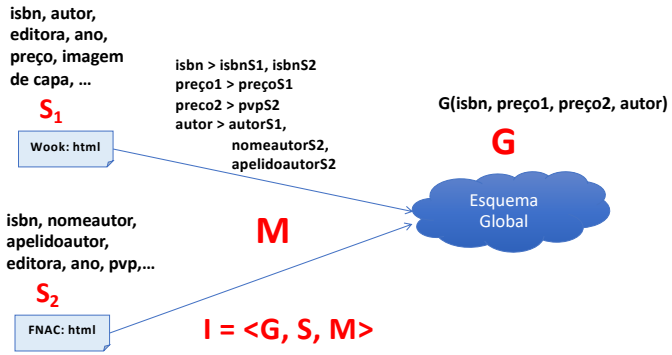
I = <G, S, M>

48

48

Integração de Dados: Wrappers - exemplo 1b

Objectivo: Inserir ISBN e obter nome do autor e preço mais baixo (duas fontes)



49

49

Integração de Dados: Wrappers

- Programas que comunicam com as fontes de dados
  - Enviam as *queries* para as fontes de dados
  - Recebem a resposta
  - Executam transformações para que a resposta seja manipulável pelo sistema

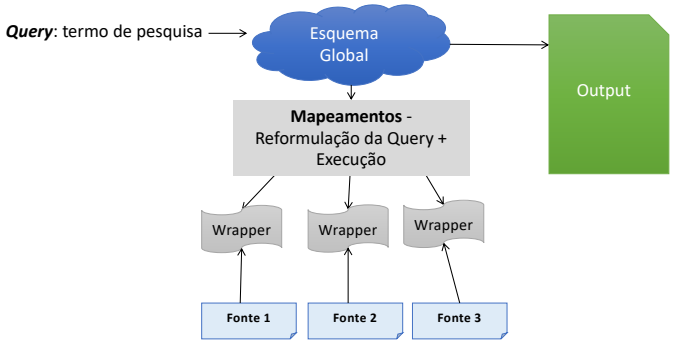
WRAPPER

- transforma uma *query* em HTTP requests
- pesquisa nas fontes (HTML, XML, TXT, ...)
- transforma a resposta (dados não estruturados) em tuplos
- devolve a informação para ser tratada pelo sistema
  - HTML, XML, TXT, ...

50

50

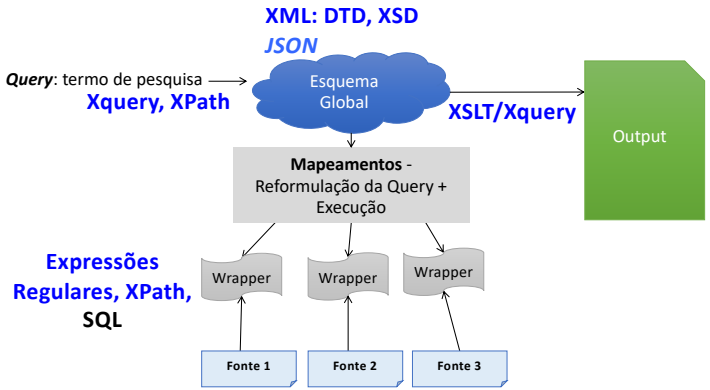
Integração de Dados: Wrappers - exemplo



51

51

Integração de Dados: Wrappers - exemplo



52

52