PRÉ PROCESSAMENTO

INTEGRAÇÃO DE DADOS

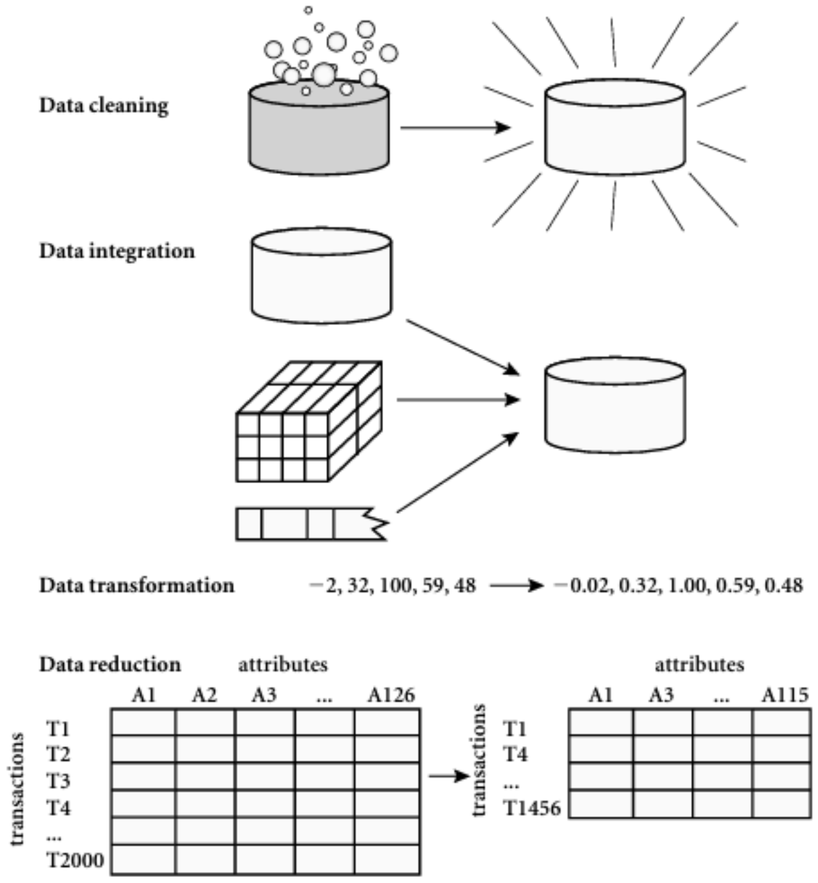
Alunos: Estevão Augusto; Gabriel Fagundes; Marco Franco, Caio, Diogo Carvalho.

Professor: Luiz Henrique de Campos Merschmann.

UNIVERSIDADE FEDERAL DE LAVRAS

NIAD

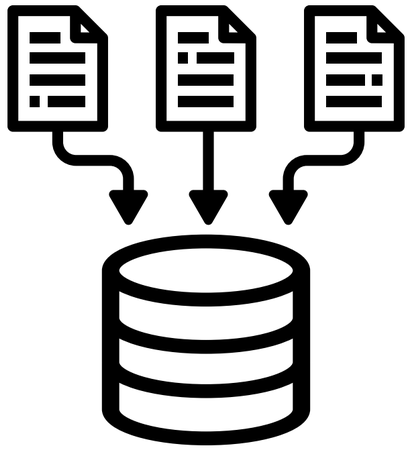
O QUE É INTEGRAÇÃO DE DADOS?



O QUE É INTEGRAÇÃO DE DADOS?

- Integração de Dados envolve a combinação de dados de múltiplas fontes em um formato de dados coerente.
  - Fontes como: bancos de dados, cubos de dados, e arquivos planos.

Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$



COM O QUE DEVO TOMAR CUIDADO?

1. Problema da Identificação de Entidades

2. Redundância de Dados

3. Conflitos de Valores de Dados

COM O QUE DEVO TOMAR CUIDADO?

1. **Problema da Identificação de Entidades**

   Como garantir que **id_cliente** de uma fonte e **num_cliente** de outra se referem à mesma pessoa?

COM O QUE DEVO TOMAR CUIDADO?

1. **Problema da Identificação de Entidades**

Metadados descrevem os dados (nome, tipo, significado, regras para nulos) e são usados para mapear corretamente atributos de fontes diferentes, evitando erros de integração.

EXEMPLOS DE METADADOS

Pesquisadores da universidade de Waikato, situada na Nova Zelândia, criaram o padrão ARFF (Formato de Arquivo Atributo-Relação) para uso com o software de pré-processamento e manipulação de dados Weka.

A diferença do ARFF para outros formatos de arquivo, como o próprio CSV, reside justamente na presença e uso de metadados. Esses metadados se encontram no cabeçalho do arquivo.

EXEMPLOS DE METADADOS

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

COM O QUE DEVO TOMAR CUIDADO?

1. **Problema da Identificação de Entidades**

2. **Redundância de Dados**

   Se uma tabela de clientes já possui a coluna "data_de_nascimento", é realmente necessário manter uma coluna para "idade"?

COM O QUE DEVO TOMAR CUIDADO?

2. Redundância de Dados

Um atributo é redundante se pode ser derivado de outros. Redundâncias também ocorrem quando nomes de atributos ou dimensões são inconsistentes, ou quando há tuplas (registros) duplicadas.

Uma forma de solucionar esse problema é com a **análise de correlação**. A análise de correlação detecta redundâncias medindo a força com que um atributo implica no outro.

COM O QUE DEVO TOMAR CUIDADO?

Para dados numéricos: Usa-se o Coeficiente de Pearson (rA,B), onde
-1 ≤ rA,B ≤ +1

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B}$$

COM O QUE DEVO TOMAR CUIDADO?

Para dados categóricos: Usa-se o Teste Qui-Quadrado ($\chi 2$). Ele testa a hipótese de que os atributos são independentes. Se a hipótese for rejeitada, os atributos são considerados estatisticamente relacionados.

COM O QUE DEVO TOMAR CUIDADO?

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

	male	*female*	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Tabela de contingência

COM O QUE DEVO TOMAR CUIDADO?

1. **Problema da Identificação de Entidades**

2. **Redundância de Dados**

3. **Conflitos de Valores de Dados**

   Se o sistema de vendas registra o peso de um produto em quilos (kg) e o sistema do fornecedor o registra em libras (lb), como garantimos que o valor é o mesmo após a integração dos dados?

COM O QUE DEVO TOMAR CUIDADO?

3. Conflitos de Valores de Dados

Causas comuns:

Diferenças de representação ou escala: Atributos podem usar unidades de medida (ex: métricas vs. imperiais) ou moedas diferentes.

Diferentes níveis de abstração: Um atributo "vendas" pode significar o total de uma filial em um sistema e o total de uma região em outro.

Estruturas de dados distintas: Um desconto pode ser aplicado ao pedido total em um sistema e por item individual em outro.

Obrigado pela atenção!