Instituto Superior de Engenharia

Politécnico de Coimbra

Integração de Dados

CTeSP Tecnologias e Programação de Sistemas de Informação (Cantanhede)

Professor: João Leal

joao.leal@isec.pt

UNIgreen The green European University

PRR Plano de Recuperação e Resiliência

REPÚBLICA PORTUGUESA

Financiado pela União Europeia NextGenerationEU

Voltando ao ETL

Exercício 1 – Detetar problemas de qualidade

Dado o *dataset clientes.csv*:

```
id,nome,email,idade
1,Ana,ana@email.com,25
2,Bruno,bruno@email,17
3,,carla@email.com,30
3,Carla,carla@email.com,30
```

1. Identificar registos com problemas (nulos, duplicados, idades inválidas, emails mal formatados).

2. Corrigir e gerar um *dataset* limpo.

Voltando ao ETL


```python
clientes = pd.read_csv("clientes.csv")

# Duplicados
clientes = clientes.drop_duplicates()

# Nome vazio → substituir por "DESCONHECIDO"
clientes["nome"].fillna("DESCONHECIDO", inplace=True)

# Corrigir emails inválidos (verificar se contém "@")
clientes = clientes[clientes["email"].str.contains("@")]

# Regra de negócio: idade >= 18
clientes = clientes[clientes["idade"] >= 18]

print(clientes)
```

Voltando ao ETL

Exercício 2 – Regras de Consistência

Dataset vendas.csv

```
id_venda,id_cliente,valor
1,10,100
2,11,200
3,99,150
```

Dataset clientes.csv

```
id_cliente,nome
10,Ana
11,Bruno
```

Problema: Existe venda com id_cliente=99, que não está em clientes.

Voltando ao ETL


```python
#ler os dois ficheiros CSV e a carregá-los em DataFrames do pandas
clientes = pd.read_csv("clientes.csv")
vendas = pd.read_csv("vendas.csv")

# Validar integridade referencial
vendas_validas = vendas[vendas["id_cliente"].isin(clientes["id_cliente"])]
print(vendas_validas)
```

Voltando ao ETL

Explicando o código…

```
#ler os dois ficheiros CSV e a carregá-los em DataFrames do pandas
clientes = pd.read_csv("clientes.csv")
vendas = pd.read_csv("vendas.csv")

# Validar integridade referencial
#Estamos a verificar se cada id_cliente presente nas vendas também existe na tabela de clientes.
#.isin(clientes["id_cliente"]) → devolve uma série de valores booleanos (True ou False) dizendo se
#o valor de id_cliente em vendas está dentro da lista de id_cliente em clientes.
vendas_validas = vendas[vendas["id_cliente"].isin(clientes["id_cliente"])]

#Mostra no ecrã apenas as vendas que têm um id_cliente válido.
#a venda com id_cliente = 99 desaparece, porque não existe esse cliente na tabela clientes.
print(vendas_validas)
```

Este código garante a integridade referencial entre as tabelas clientes e vendas → só ficam registadas as vendas associadas a clientes que realmente existem.