



# Licenciatura em Engenharia de Sistemas Informáticos Integração de Sistemas de Informação 2025/2026

Análise de Dados Meteorológicos

João Carlos da Silva Lopes Nº 12168 - LESI PL

**Docente: Luís Ferreira** 

# Índice

Glossário	4
Introdução	5
Contexto	6
Objetivos	7
Estratégia	8
Dados	9
Transformação 1	11
Jobs	14
Transformação 2	17
Power BI	19
Conclusão	21
Trabalhos Futuros	22
Bibliografia	23

# Índice de Figuras

Figura 1 - Variável que guarda os dados no ficheiro CSV	9
Figura 2 - Ficheiros carregados no Pentaho	9
Figura 3 - Dados do ficheiro csv	10
Figura 4 - Transformação	
Figura 5 - Expressão Regular	11
Figura 6 - Novo campo e distinção da data	12
Figura 7 - Linha que foram para os ficheiros	12
Figura 8 - Validação	12
Figura 9 - Linhas apenas datas válidas	
Figura 10 - Exemplo output do ficheiro xml	13
Figura 11 - job para gerar texto	14
Figura 12 - Texto gerado a partir do job	14
Figura 13 - Job completo	14
Figura 14 - Confirmação	
Figura 15 - Caixa de entrada do email	15
Figura 16 - Texto do email	15
Figura 17 - Configuração	15
Figura 18 - Configuração	
Figura 19 - Configuração	
Figura 20 - Transformação	17
Figura 21 - Dados do ficheiro Clientes	17
Figura 22 - Dados do ficheiro Vendas	17
Figura 23 - Fórmulas aplicadas	18
Figura 24 - Ficheiro gerado em xml	
Figura 25 - Dados retirados do xml alterado	19
Figura 26 - Gráfico com as temperaturas registadas	
Figura 27 - Cidades e temperaturas	20

#### Glossário

- **ETL:** Extract, Transform, Load. Processo que extrai dados de várias fontes, transforma-os para garantir consistência e carrega-os num destino final para análise ou armazenamento.
- **Pentaho:** Ferramenta de software usada para criar e automatizar processos ETL, permitindo extrair, transformar e carregar dados de forma visual e integrada.
- **Power BI:** ferramenta da Microsoft para visualização e análise de dados, que transforma informação bruta em dashboards e relatórios interativos.
- **Dashboard:** painel visual que apresenta dados de forma interativa através de gráficos, tabelas e indicadores para facilitar a análise e a tomada de decisões.
- **API:** Application Programming Interface. Conjunto de regras que permite a comunicação e troca de dados entre diferentes aplicações ou sistemas.
- **CSV:** Comma-Separated Values. Formato de ficheiro de texto simples que armazena dados em forma de tabela, separados por vírgulas ou ponto e vírgula.
- TXT: formato de ficheiro de texto simples que contém dados sem formatação, usado para armazenar informação de forma legível e leve.
- **JSON:** JavaScript Object Notation. Formato de dados leve e estruturado, usado para armazenar e trocar informação entre sistemas de forma simples e legível.
- XML: Extensible Markup Language. Formato de marcação que organiza dados em estruturas hierárquicas através de etiquetas, facilitando a troca de informação entre sistemas.
- **Transformação:** processo no Pentaho que define as etapas de extração, tratamento e exportação de dados, compondo o núcleo do fluxo ETL.
- **Job:** processo de controlo no Pentaho que executa uma ou várias transformações e tarefas de forma automatizada e sequencial.
- **Expressão Regular:** padrão textual usado para pesquisar, validar ou extrair partes de strings com base em regras de correspondência.

#### Introdução

O presente projeto de avaliação faz parte da Unidade Curricular de Integração de Sistemas de Informação lecionada no 1º semestre do 3º Ano do curso Engenharia de Sistemas Informáticos, do Instituto Politécnico do Cávado e do Ave.

O intuito deste trabalho prático é aprender a aplicar e experimentar ferramentas recorrendo a processos ETL inerentes a processos de ISI ao nível dos dados, com a utilização de scripts próprias. Neste trabalho irei utilizar a ferramenta Pentaho.

Um ETL é responsável pela integração de várias etapas: extração, transformação e carregamento de dados, aplicando boas práticas de tratamento, validação e automação. Como resultado é esperado que a informação saia limpa e estruturada para posterior análise.

Na etapa final é suposto que os resultados obtidos sejam integrados num Dashboard desenvolvido no Power BI Desktop permitindo a visualização e análise dos dados processados.

Esta junção entre Pentaho e Power Bi reforça o conceito de um pipeline de dados completo, onde o processamento técnico e a análise visual se complementam para fornecer informação estruturada e de apoio à decisão.

Assim, este trabalho não só ilustra as capacidades técnicas do Pentaho Data Integration no contexto de integração e automação de dados, como também evidencia a importância da visualização analítica na fase final do processo, oferecendo uma perspetiva prática e aplicada sobre os conceitos fundamentais de sistemas de informação empresariais.

#### **Contexto**

Os processos de negócio não param de se reconfigurar, sendo que as empresas enfrentam desafios diários e constantes de análise e aquisição de novas soluções informáticos que facilitem o seu trabalho.

Existe uma enorme necessidade de rentabilizar a aquisição de soluções, tanto a nível do investimento financeiro bem como a nível da dependência dos processos, dos dados que deles fazem parte. Para isto acontecer as empresas procuram formas capazes de analisarem mais-valias e menos impactos com as novas aquisições.

Processos como: auditorias a dados, processos, segurança; migração e reorganização de dados; análise e processamento de dados; coleta de dados contínuos oriundos de infraestruturas da Internet, tais como sensores em máquinas de linhas industriais; recomendações e previsões sobre estados com processamento recorrente a big data e mesmo inteligência artificial, são exemplos de cenários onde processos de ETL poderão desempenhar papel preponderante.

No contexto emergente dos smart environments (Cities, Health Care, Social, etc.) a evidência deste tipo de soluções é clara e crescente. A integração de soluções ditas mais inteligentes, que reagem a eventos, em processos existentes ou mesmo legados, são desafios constantes.

Pretende-se então que com a execução de processos ETL se consiga resolver parte destas dificuldades encontradas pelas empresas com vista à otimização dos processos já desenvolvidos nas mesmas.

# **Objetivos**

Os objetivos do desenvolvimento deste trabalho são os seguintes:

- Consolidar conceitos associados à Integração de Sistemas de Informação usando Dados;
- Analisar e especificar cenários de aplicação de processos de ETL;
- Explorar ferramentas de suporte a processos de ETL;
- Explorar novas Tecnologias, Frameworks ou Paradigmas;
- Potenciar a experiência no desenvolvimento de software;
- Facilitar a assimilação do conteúdo da Unidade Curricular.
- Contribuir para qualidade da escrita de relatórios

#### Estratégia

A estratégia adotada no desenvolvimento deste trabalho baseou-se numa abordagem prática e iterativa permitindo construir progressivamente todas as fases de um processo ETL e garantir a sua integração no final, dando resposta a todos os pontos pedidos no enunciado.

Para o planeamento foram definidos os objetivos principais. Criar uma transformação capaz de importar dados, validar a sua consistência, aplicar operações de transformação e exportar resultados em diferentes formatos.

Para a importação foram carregados vários formatos de ficheiros, csv, xml, txt, json de forma a tratar os seus dados. Foi utilizado um ficheiro csv onde se recolheu a informação e analisou-se separando a informação em registos válidos e inválidos. Foram aplicados componentes mais avançados tais como GroupBy, LoorkUps, Calculator e posterior exportação da informação.

Na parte da automação houve a criação de um job incorporando várias etapas onde se utilizou a funcionalidade do Mail.

Para demonstrar alterações nos dados foi criada uma de análise no power BI Besktop, alimentada pelos ficheiros gerados pelo Pentaho.

Esta abordagem permite uma integração completa entre a extração técnica e a interpretação visual dos resultados, proporcionando uma visão global sobre o ciclo de vida dos dados.

Assim, a estratégia seguida privilegiou a organização, modularidade e validação contínua, garantindo um desenvolvimento controlado.

#### **Dados**

Os dados utilizados no trabalho estão divididos em duas pastas.

- Entrada: ficheiros gerados em código c# de onde foram retirados os dados.
- Saída: ficheiros gerados a partir do Pentaho.

Nestas duas pastas existem ficheiros com vários formatos: csv, json, txt e xml.



Figura 1 - Variável que guarda os dados no ficheiro CSV

As restantes variáveis responsáveis por guardar os dados dos restantes ficheiros, têm todas a mesma forma.

Os ficheiros criados com recurso ao C# foram então carregados no Pentaho:



Figura 2 - Ficheiros carregados no Pentaho

Exemplo de um ficheiro csv carregado no Pentaho e a forma de como os dados aparecem demonstrados na ferramenta.

Rows of step: teste_csv (18 rows)							
#	data_hora	local	temp_c	humidade_per	vento_kmh	lat_c	long_c
1	01-10-2025 00:00	Aveiro	10.1	10	10.1	10.1	-38
2	01-10-2025 01:00	Beja	11.2	11	11.2	11.2	-26.7
3	01-10-2025 02:00	Braga	12.3	12	12.3	12.3	-25.6
4	02-10-2025 03:00	Braganca	13.4	13	13.4	13.4	-24.5
5	02-10-2025 04:00	C.Branco	14.5	14	14.5	14.5	-23.4
6	02-10-2025 05:00	Coimbra	15.6	15	15.6	15.6	-22.3
7	03-10-2025 06:00	Evora	16.7	16	16.7	16.7	-21.2
8	03-10-2025 07:00	Faro	17.8	17	17.8	17.8	-20.1
9	03-10-2025 08:00	Guarda	18.9	18	18.9	18.9	-19
10	04-10-2025 09:00	Leiria	19	19	19	19	-18.9
11	04-10-2025 10:00	Lisboa	20.1	20	20.1	20.1	-17.8
12	04-10-2025 11:00	Portalegre	21.2	21	21.2	21.2	-16.7
13	05-10-2025 12:00	Porto	22.3	22	22.3	22.3	-15.6
14	05-10-2025 13:00	Santarem	23.4	23	23.4	23.4	-14.5
15	05-10-2025 14:00	Setubal	24.5	24	24.5	24.5	-13.4
16	06-10-2025 15:00	V.Castelo	25.6	25	25.6	25.6	-12.3
17	06-10-2025 16:00	V.Real	26.7	26	26.7	26.7	-11.2
18	06-10-2025 17:00	Viseu	27.8	27	27.8	27.8	-10.1

Figura 3 - Dados do ficheiro csv

## Transformação 1

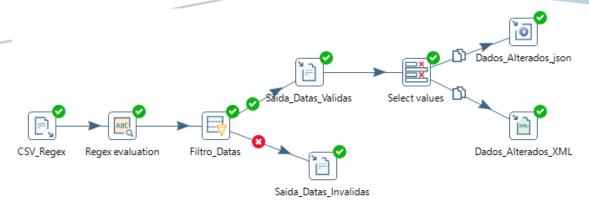


Figura 4 - Transformação

Esta transformação lê um ficheiro csv que contém dados meteorológicos, verifica se os campos da data estão corretos, separa as linhas válidas e inválidas e exporta essa informação para 2 ficheiros.

- É adicionado um ficheiro csv no Pentaho com os dados meteorológicos, com vários campos preenchidos.
- Cada linha do ficheiro representa um registo: data\_hora, local, temp\_c, humidade\_per, vento\_kmh, lat\_c, long\_c.
- É aplicada ao ficheiro uma expressão regular para validar o formato do campo data\_hora.

Com esta expressão pretendemos validar o dia para valores entre 01/31, o mês para valores entre 01/12 e o ano é aceite se começar com 19/20.

A expressão verifica ainda que a hora está compreendida entre 00/23 e os minutos entre 00/59.

#### Regular expression:

^(0[1-9]|[12][0-9]|3[01])-(0[1-9]|1[0-2])-(19|20)\d{2}\s([01][0-9]|2[0-3]):([0-5][0-9])\$

Figura 5 - Expressão Regular

O objetivo é garantir que a data segue o padrão correto, neste caso: dd\_mm\_yyyy
 hh:mm. Foi criado um campo novo (data\_valida) de forma a verificar quais datas
 é que cumprem com os critérios

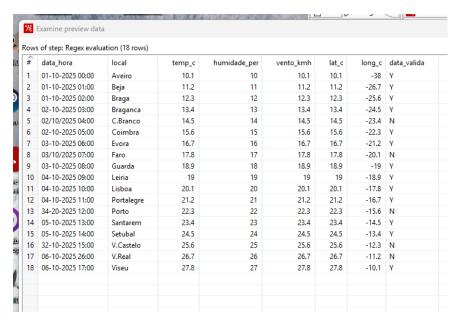


Figura 6 - Novo campo e distinção da data

• As linhas com as datas foram filtradas e colocadas em 2 ficheiros distintos.

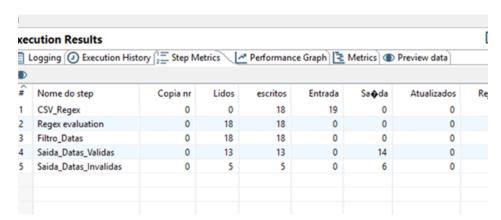


Figura 7 - Linha que foram para os ficheiros



Figura 8 - Validação

- Esta triagem permite identificar os erros nos dados de origem.
- Depois foi selecionado o ficheiro com as datas válidas e apenas foram extraídos os dados que considerei importantes e também mudados os seus nomes. Os novos campos definidos para serem exportados foram: data\_e\_hora, localidade e temperatura.

Ŷ	data_e_hora	localidade	temperatura
	01-10-2025 00:00	Aveiro	10.1
2	01-10-2025 01:00	Beja	11.2
3	01-10-2025 02:00	Braga	12.3
ļ	02-10-2025 03:00	Braganca	13.4
5	02-10-2025 05:00	Coimbra	15.6
5	03-10-2025 06:00	Evora	16.7
7	03-10-2025 08:00	Guarda	18.9
3	04-10-2025 09:00	Leiria	19
9	04-10-2025 10:00	Lisboa	20.1
10	04-10-2025 11:00	Portalegre	21.2
11	05-10-2025 13:00	Santarem	23.4
2	05-10-2025 14:00	Setubal	24.5
3	06-10-2025 17:00	Viseu	27.8

Figura 9 - Linhas apenas datas válidas

 Foram então os dados exportados para 2 ficheiros diferentes, um ficheiro xml e outro ficheiro json.

Figura 10 - Exemplo output do ficheiro xml

#### **Jobs**

Para a criação do job foi utilizada uma transformação bastante simples apenas para gerar um ficheiro com alguma informação.



Figura 11 - job para gerar texto

1;Teste;Aveiro;20.5 2;Teste;Aveiro;20.5 3;Teste;Aveiro;20.5

Figura 12 - Texto gerado a partir do job

Esta transformação foi realizada para ser aproveitada no job que foi feita a seguir.



Figura 13 - Job completo

A job tem um ponto de partida: Start. O processo é iniciado nesta etapa.

- Após este passo crio uma pasta para onde os ficheiros de saída serão gravados.
- Depois é aplicada a transformação apresentada anteriormente.
- É aplicado um log para informar que a transformação foi realizada com sucesso.
- É enviado um email a informar que a transformação ocorreu com sucesso.
- E a etapa success marca o fim do processo, informando que todas as etapas foram realizadas sem falhas.

```
025-10-08 03:14:20.002 - Job_Exportar_Dados - Starting entry [Write to log]
025-10-08 03:14:20.003 - - Exportação concluída com sucesso! Ficheiro gerado na pasta pedida.
025-10-08 03:14:20.004 - Job_Exportar_Dados - Starting entry [Mail]
025-10-08 03:14:31.905 - Job_Exportar_Dados - Starting entry [Success]
025-10-08 03:14:31.909 - Job_Exportar_Dados - Finished job entry [Success] (result=[true])
025-10-08 03:14:31.910 - Job_Exportar_Dados - Finished job entry [Mail] (result=[true])
025-10-08 03:14:31.910 - Job_Exportar_Dados - Finished job entry [Write to log] (result=[true])
025-10-08 03:14:31.911 - Job_Exportar_Dados - Finished job entry [Transformation] (result=[true])
025-10-08 03:14:31.911 - Job_Exportar_Dados - Finished job entry [Create a folder] (result=[true])
025-10-08 03:14:31.912 - Job_Exportar_Dados - Job execution finished
```

Figura 14 - Confirmação

□ ☆ eu Teste Pentaho - Exportação realizada com sucesso. Job: ----- JobName : Job\_Exportar\_Dados Directory : / JobEntry : Mail Previous results: ----------- Job entry Nr : 2 Errors : 0 Lin...

Figura 15 - Caixa de entrada do email

#### Teste Pentaho Caixa de entrada x

Teste <joaocarlosafb10@gmail.com> para mim ▼

Exportação realizada com sucesso.

Figura 16 - Texto do email

Para esta etapa do Mail foi necessário haver uma configuração.

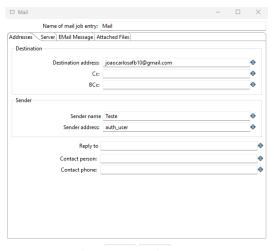


Figura 17 - Configuração

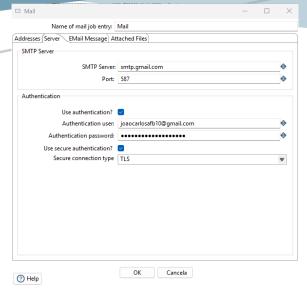


Figura 18 - Configuração

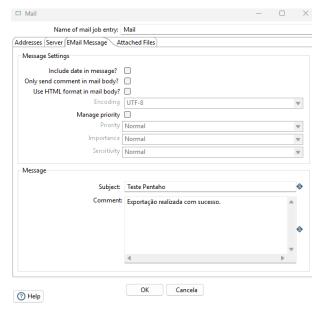


Figura 19 - Configuração

## Transformação 2

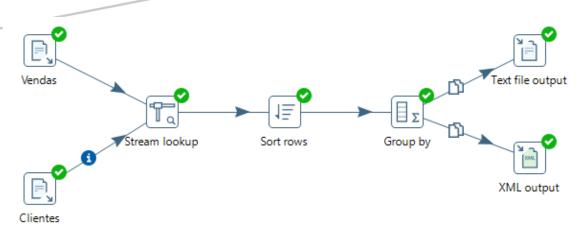


Figura 20 - Transformação

Esta transformação junta informações de dois ficheiros: Clientes e Vendas, ordena os dados, calcula totais agrupados por cliente, e exporta os resultados em dois formatos de saída: .txt e .xml.

#### Rows of step: Vendas (5 rows)

#	ID_Venda	ID_Cliente	Valor
1	100	1	10
2	101	2	20
3	102	3	30
4	103	4	40
5	104	5	50

Figura 22 - Dados do ficheiro Vendas

#### Rows of step: Clientes (10 rows)

#	ID_Cliente	Nome
1	1	João
2	2	Rodrigo
3	3	Joana
4	4	Martinho
5	5	Fernando
6	6	Joana
7	7	Rebelo
8	8	Tom
9	9	Antonio
10	10	Marcelo

Figura 21 - Dados do ficheiro Clientes

- No ID\_Cliente deve ler-se: ID\_Vendedor e o ficheiro Vendedores.
- Lê os dados dos dois ficheiros.
- Faz uma junção entre os dois fluxos de dados. O objetivo é combinar a informação do cliente com as respetivas vendas, com base no ID\_Cliente.
- É aplicado uma ordenação das linhas pelo Nome Cliente.

Agrupa os registos de acordo com as operações que se quiser relacionar.

#### Aggregates:

#	Name	Subject	Туре
1	Total_Vendas	Valor	Sum
2	Media_Vendas	Valor	Average (Mean)
3	Qtd_Vendas	Valor	Maximum

Figura 23 - Fórmulas aplicadas

 Foram filtradas as informações relacionadas com o total de vendas, com a media das vendas e com a quantidade vendida. Informação essa que foi extraída para um ficheiro .txt e outro .xml.

```
▼<Rows>
 ▼ < Row>
    <nome>Fernando</nome>
    <total_vendas>50</total_vendas>
    <media_vendas>50</media_vendas>
    <qts_vendas>50</qts_vendas>
   </Row>
 ▼<Row>
     <nome>Joana</nome>
     <total_vendas>30</total_vendas>
     <media_vendas>30</media_vendas>
     <qts_vendas>30</qts_vendas>
     <nome>João</nome>
     <total_vendas>10</total_vendas>
     <media_vendas>10</media_vendas>
     <qts_vendas>10</qts_vendas>
     <nome>Martinho</nome>
     <total_vendas>40</total_vendas>
     <media_vendas>40</media_vendas>
     <qts_vendas>40</qts_vendas>
   </Row>
     <nome>Rodrigo</nome>
     <total_vendas>20</total_vendas>
     <media_vendas>20</media_vendas>
     <qts_vendas>20</qts_vendas>
   </Row>
 </Rows>
```

Figura 24 - Ficheiro gerado em xml

#### **Power BI**

- Importação de um ficheiro em formato xml.
- Limpeza dos dados e transformação no Power Query, trocado os pontos pelas virgulas para melhor interpretação por parte do programa.
- Criação de dois gráficos: temperaturas medidas ao longo de um intervalo de dias e ainda os números de cidades que atingem determinada temperatura.

O resultado final foi um dashboard interativo que permite analisar de forma visual e intuitiva os dados processados na transformação.

#### Row

data_e_hora	localidade	temperatura
01/10/2025 00:00:00	Aveiro	10.1
01/10/2025 01:00:00	Beja	11.2
01/10/2025 02:00:00	Braga	12.3
02/10/2025 03:00:00	Braganca	13.4
02/10/2025 05:00:00	Coimbra	15.6
03/10/2025 06:00:00	Evora	16.7
03/10/2025 08:00:00	Guarda	18.9
04/10/2025 09:00:00	Leiria	19
04/10/2025 10:00:00	Lisboa	20.1
04/10/2025 11:00:00	Portalegre	21.2
05/10/2025 13:00:00	Santarem	23.4
05/10/2025 14:00:00	Setubal	24.5
06/10/2025 17:00:00	Viseu	27.8

Figura 25 - Dados retirados do xml alterado



Figura 26 - Gráfico com as temperaturas registadas

Este gráfico mostra as temperaturas que foram medidas ao longo dos dias, dia 1 ao dia 6 de outubro de 2025.

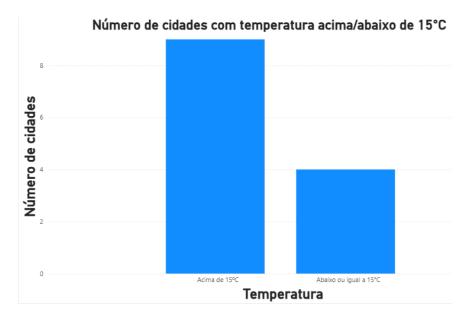


Figura 27 - Cidades e temperaturas

Este gráfico mostra o número de cidades onde foram medidas temperatura acima abaixo e acima dos 15ºC.

## Conclusão

Este trabalho permitiu compreender de forma prática como funciona um processo ETL e como as várias etapas - extração, transformação e carga - se interligam dentro do Pentaho Data Integration (Spoon).

Foi possível criar uma transformação completa que lê dados, valida e trata a informação, e exporta os resultados em diferentes formatos (CSV, JSON e XML). A criação de um Job automatizado permitiu ainda controlar todo o processo, gerar logs e enviar notificações por e-mail, tornando o sistema mais fiável e profissional.

Por fim, a ligação ao Power BI mostrou a importância da análise visual dos dados, ajudando a transformar os resultados do ETL em informação útil.

No geral, o trabalho foi bastante enriquecedor, pois juntou a parte técnica da integração de dados com a parte analítica, permitindo ter uma visão completa do ciclo de vida da informação.

## **Trabalhos Futuros**

Tendo em conta o pouco tempo que tive para realizar este trabalho, pretendo melhorar o mesmo e implementar todas as funcionalidades que faltam relativamente ao que foi pedido.

De resto, existe sempre espaço para melhorar e para evoluir.

# **Bibliografia**

- https://elearning.ipca.pt/2526/course/view.php?id=38561
- https://pentaho.com/pentaho-ee-onprem/
- <a href="https://www.microsoft.com/pt-pt/power-platform/products/power-bi?market=pt">https://www.microsoft.com/pt-pt/power-platform/products/power-bi?market=pt</a>
- $\ https://support.microsoft.com/pt-pt/office/xml-para-principiantes-a 87d 234d-4c 2e-4409-9cbc-45e 4eb 857d 44$