****

**Licenciatura em Engenharia de Sistemas Informáticos**

**Integração de Sistemas de Informação**

**2025/2026**

**Análise de Dados Musicais**

**Miguel Silva Areal**

**Nº 29559 – LESI PL**

**Docente: Luís Ferreira**

**Índice**

[**Glossário** 4](#_Toc210858066)

[**Introdução** 5](#_Toc210858067)

[**Contexto** 6](#_Toc210858068)

[**Objetivos** 7](#_Toc210858069)

[**Estratégia** 8](#_Toc210858070)

[**Dados** 9](#_Toc210858071)

[**Transformação 1** 11](#_Toc210858072)

[**Jobs** 14](#_Toc210858073)

[**Transformação 2** 17](#_Toc210858074)

[**Power BI** 19](#_Toc210858075)

[**Conclusão** 21](#_Toc210858076)

[**Trabalhos Futuros** 22](#_Toc210858077)

[**Bibliografia** 23](#_Toc210858078)

**Índice de Figuras**

[Figura 1 - Variável que guarda os dados no ficheiro CSV 9](#_Toc210858039)

[Figura 2 - Ficheiros carregados no Pentaho 9](#_Toc210858040)

[Figura 3 - Dados do ficheiro csv 10](#_Toc210858041)

[Figura 4 - Transformação 11](#_Toc210858042)

[Figura 5 - Expressão Regular 11](#_Toc210858043)

[Figura 6 - Novo campo e distinção da data 12](#_Toc210858044)

[Figura 7 - Linha que foram para os ficheiros 12](#_Toc210858045)

[Figura 8 - Validação 12](#_Toc210858046)

[Figura 9 - Linhas apenas datas válidas 13](#_Toc210858047)

[Figura 10 - Exemplo output do ficheiro xml 13](#_Toc210858048)

[Figura 11 - job para gerar texto 14](#_Toc210858049)

[Figura 12 - Texto gerado a partir do job 14](#_Toc210858050)

[Figura 13 - Job completo 14](#_Toc210858051)

[Figura 14 - Confirmação 15](#_Toc210858052)

[Figura 15 - Caixa de entrada do email 15](#_Toc210858053)

[Figura 16 - Texto do email 15](#_Toc210858054)

[Figura 17 - Configuração 15](#_Toc210858055)

[Figura 18 - Configuração 16](#_Toc210858056)

[Figura 19 - Configuração 16](#_Toc210858057)

[Figura 20 - Transformação 17](#_Toc210858058)

[Figura 21 - Dados do ficheiro Clientes 17](#_Toc210858059)

[Figura 22 - Dados do ficheiro Vendas 17](#_Toc210858060)

[Figura 23 - Fórmulas aplicadas 18](#_Toc210858061)

[Figura 24 - Ficheiro gerado em xml 18](#_Toc210858062)

[Figura 25 - Dados retirados do xml alterado 19](#_Toc210858063)

[Figura 26 - Gráfico com as temperaturas registadas 20](#_Toc210858064)

[Figura 27 - Cidades e temperaturas 20](#_Toc210858065)

# **Glossário**

**• ETL (Extract, Transform, Load):** Processo que extrai dados de várias fontes, transforma-os para garantir consistência e qualidade, e carrega-os num destino final para análise ou armazenamento.

**• KNIME:** Ferramenta open-source para criação e automação de processos ETL. Permite extrair, transformar e carregar dados de forma visual e integrada, através de fluxos compostos por nós (nodes).

**• Node-RED:** Ferramenta open-source desenvolvida pela IBM usada para integrar sistemas e fluxos de dados através de uma interface visual baseada em nós (nodes). Utilizada frequentemente em IoT, automação e comunicação entre APIs e dispositivos.

**• Dashboard:** Painel visual que apresenta dados de forma interativa, através de gráficos, tabelas e indicadores, facilitando a análise e a tomada de decisões.

**• API (Application Programming Interface):** Conjunto de regras e definições 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 formato tabular, separados por vírgulas ou ponto e vírgula.

**• SQL (Structured Query Language):** Linguagem de consulta estruturada usada para gerir e manipular bases de dados relacionais.

**• JSON (JavaScript Object Notation):** Formato leve e estruturado para armazenar e trocar dados entre sistemas, de forma simples e legível por humanos e máquinas.

**• SQLite:** Sistema de gestão de bases de dados relacional leve e embutido, que armazena dados em ficheiros locais. É amplamente usado em aplicações que necessitam de bases de dados simples e portáteis.

**• DB Browser for SQLite:** Aplicação gráfica que permite criar, visualizar, editar e gerir bases de dados SQLite de forma simples e intuitiva.

**• Transformação:** Conjunto de operações aplicadas aos dados dentro de um fluxo (workflow), como limpeza, filtragem, junção e cálculo de novos campos. Representa o núcleo do processo ETL.

* **Expressão Regular:** padrão textual usado para pesquisar, validar ou extrair partes de strings com base em regras de correspondência.
* **Dataset:** Coleção de dados relacionados, geralmente organizada de forma estruturada (como num ficheiro JSON ou ficheiro CSV), que pode conter vários tipos de informação como números, texto, imagens ou áudio.
* **Access Token:**
* GET Requests:
* Endpoint:

# **Introdução**

O presente projeto de avaliação integra a Unidade Curricular de Integração de Sistemas de Informação, lecionada no 1.º semestre do 3.º ano do curso de Engenharia de Sistemas Informáticos do Instituto Politécnico do Cávado e do Ave.

O principal objetivo deste trabalho prático é aplicar e explorar ferramentas associadas a processos ETL, no âmbito dos Sistemas de Informação, recorrendo à utilização de scripts próprias.

Um processo ETL (Extract, Transform, Load) é responsável pela integração de três etapas fundamentais: Extração, transformação e carregamento de dados, garantindo a aplicação de boas práticas de tratamento, validação e automação. O resultado esperado é a obtenção de informação limpa, estruturada e consistente, adequada para posterior análise.

Na fase final, pretende-se que os resultados obtidos sejam integrados num dashboard desenvolvido no Node-RED, possibilitando a visualização e análise interativa dos dados processados.

A integração entre KNIME, SQL e Node-RED demonstra a construção de um pipeline de dados completo, no qual o processamento técnico e a análise visual se complementam, permitindo gerar informação estruturada e útil para a tomada de decisão.

Desta forma, o projeto evidencia não só as capacidades técnicas do KNIME na integração e automação de dados, mas também a relevância da visualização analítica na fase final do processo, proporcionando uma abordagem prática e aplicada aos conceitos fundamentais dos sistemas de informação empresariais.

# **Contexto**

No atual panorama empresarial, os processos de negócio encontram-se em constante transformação. As organizações são diariamente confrontadas com desafios que exigem capacidade de adaptação, análise e integração de novas soluções informáticas que otimizem o seu funcionamento. A transformação digital, aliada à crescente competitividade do mercado, obriga as empresas a tomarem decisões estratégicas fundamentadas em dados e em ferramentas tecnológicas que potenciem a sua eficiência e rentabilidade.

Neste cenário, torna-se essencial maximizar o retorno do investimento em soluções informáticas, garantindo simultaneamente a coerência e continuidade dos processos de negócio e a integridade dos dados que os sustentam. A escolha de novas tecnologias deve, por isso, ser cuidadosamente avaliada, considerando não só os benefícios diretos que proporciona, mas também os impactos que pode gerar a nível operacional, financeiro e organizacional. Assim, as empresas procuram metodologias e ferramentas que lhes permitam analisar de forma rigorosa as vantagens e desvantagens associadas a cada nova aquisição tecnológica, reduzindo riscos e potenciando oportunidades.

Com este enquadramento em mente, foi concebida uma solução que visa responder a um problema simulado, mas inspirado em situações reais do contexto empresarial:

**Problema:**  
Uma editora discográfica pretende compreender melhor o panorama musical atual e identificar as oportunidades de investimento mais promissoras. O seu objetivo é determinar com quais artistas, géneros ou estilos musicais deve estabelecer parcerias para alcançar maior sucesso económico nos serviços de streaming.

**Solução:**  
Recorrendo a dados musicais do mercado, bem como a ferramentas de tratamento, análise e visualização de dados, será desenvolvido um estudo que permita identificar padrões de consumo, tendências emergentes e fatores determinantes para o sucesso no setor musical. Esta análise apoiará a tomada de decisão estratégica da editora, fornecendo-lhe uma base sólida para direcionar os seus investimentos e colaborações futuras de forma informada e sustentada.

# **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 foi carregado um *dataset* de dados de músicas do serviço de streaming de músicas *Spotify* em formato *CSV* como a base dos dados desta análise.

A partir da informação do *dataset* foi recolhida e agregada informação adicional proveniente da *API* do *Spotify*, como dados extra de artistas e músicas.

Toda a informação disponível sofreu transformações e separada de forma a ser inserida numa base de dados SQLite.

Para demonstrar alterações nos dados foi criada uma de análise no Node-Red, alimentada pela base de dados criada.

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 projeto estão divididos em diferentes pastas.

Figura - Estrutura hierárquica dos dados utilizados.

Legenda:

* Data/Input/spotify\_songs.csv
  + *Dataset* de entrada que contém cerca de 30000 músicas, no formato *CSV*.
* Data/Output/musicas\_encontradas.csv
  + Ficheiro *CSV* que contém todas as músicas encontradas durante o processo ETL, é enviado como anexo no e-mail de notificação final.
* Data/Output/TP1-ISI-Spotify.db
  + Ficheiro de base de dados *SQL* que contém todos os dados da transformação ETL, separado em diferentes tabelas.
  + Extraído a partir de <https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs>
* Data/Keys/spotify\_keys.json
  + Ficheiro em formato *JSON* que contém as credenciais de acesso à *API* do *Spotify*, estas sendo o client\_id e client\_secret.

# **Processo ETL**

## Autenticação na API do Spotify

Antes de dar início à transformação ETL dos dados, são criadas as condições necessárias para a autenticação na *API* do Spotify.



Figura – Workflow de autenticação no serviço do Spotify

Neste processo, o objetivo é obter um *Access Token.* Este é responsável por nos permitir executar todas as GET Requests necessárias para podermos completar os dados do dataset original.

Para isto, precisamos de primeiramente ter uma conta Spotify, e criar uma app através do painel de desenvolvedor (<https://developer.spotify.com/>). Esta app permite-nos requisitar as credenciais *client\_id* e *client\_secret*, que neste contexto ficam gravadas em “data/keys/spotify\_keys.json” no formato:

Figura - Formato exemplo do ficheiro de dados de credenciais.

{

    "client\_id":"12345",

    "client\_secret":"12345"

}

Este ficheiro é importado e lido pelo JSON Reader, e os seus dados são extraídos para colunas utilizando o JSON Path. De seguida, estas novas colunas são transformadas em variáveis do workflow para que possam ser convertidas para o tipo de dados *Credential.* Esta *credential* pode então ser utilizada facilmente pelo node de autenticação, que faz um pedido de requisição do *Access* *Token* pelo *Endpoint* <https://accounts.spotify.com/api/token>. Se aceite, podemos agora utilizar este node para todos os próximos pedidos á *API* do *Spotify.*

## Leitura de Dataset



Figura - Workflow de Leitura de dataset e início de ciclo.

Neste processo, é utilizado o node CSV Reader para efetuar a leitura do dataset que contém cerca de 30000 dados de músicas

# **Transformação 1**

Figura - 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.

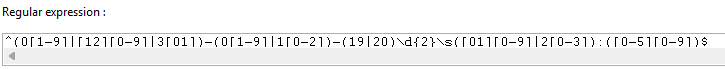


Figura - Expressão Regular

* 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

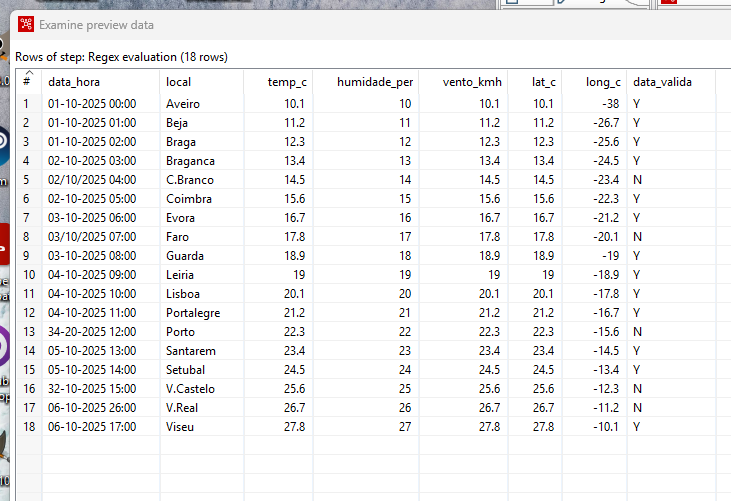


Figura - Novo campo e distinção da data

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

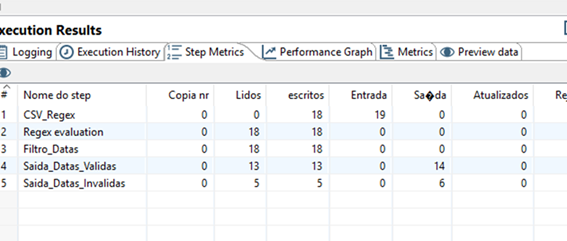


Figura - Linha que foram para os ficheiros

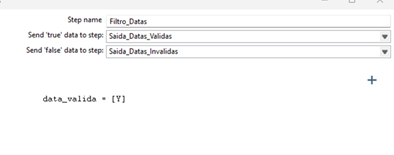


Figura - Validação

* 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.

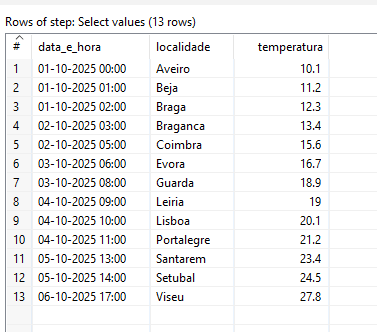


Figura - Linhas apenas datas válidas

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

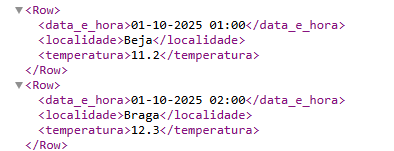


Figura - Exemplo output do ficheiro xml

# **Jobs**

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



Figura - job para gerar texto



Figura - Texto gerado a partir do job

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

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

Figura - Job completo

* 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.

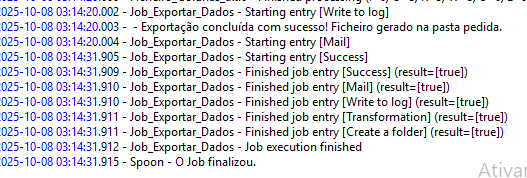


Figura - Confirmação



Figura - Caixa de entrada do email



Figura - Texto do email

Para esta etapa do Mail foi necessário haver uma configuração.

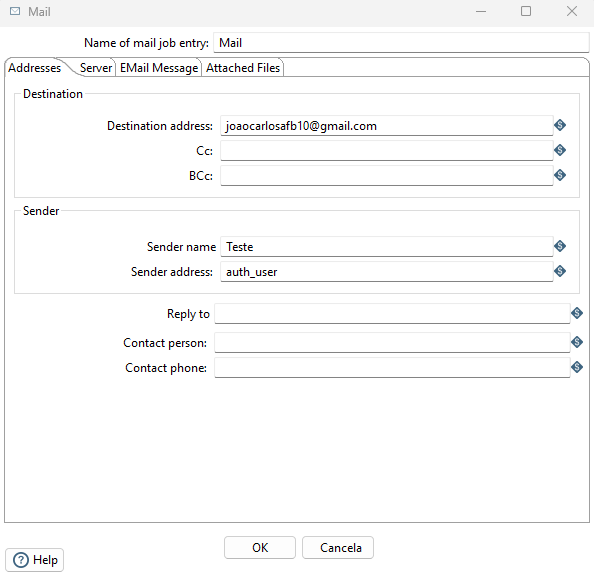


Figura - Configuração

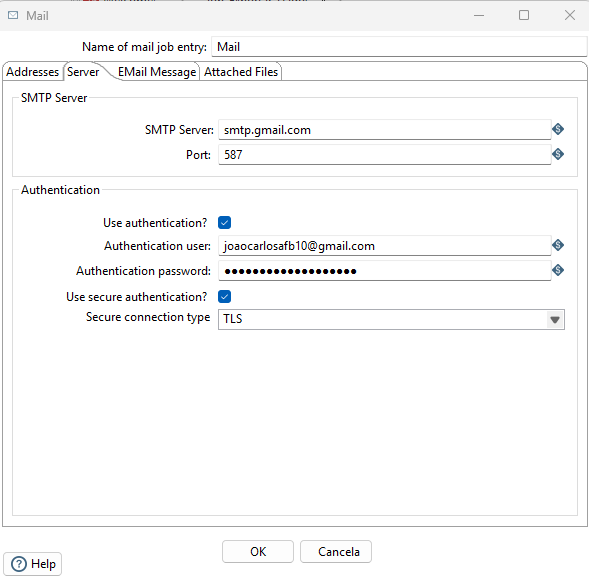


Figura - Configuração

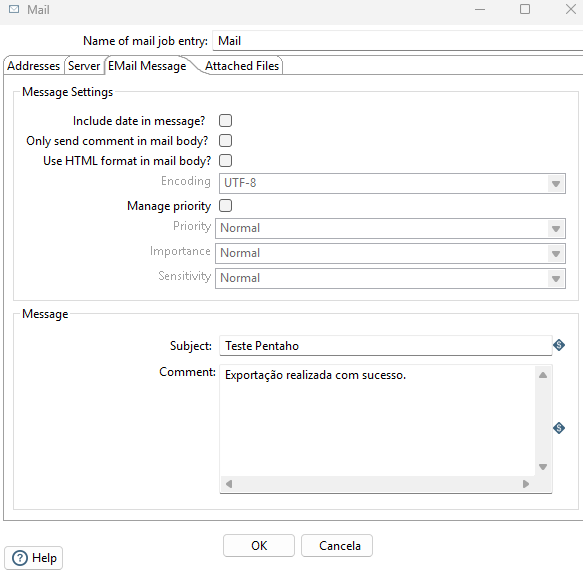


Figura - Configuração

# **Transformação 2**

Figura - 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.

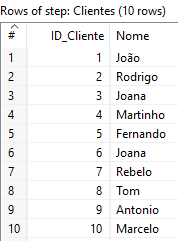


Figura - Dados do ficheiro Clientes

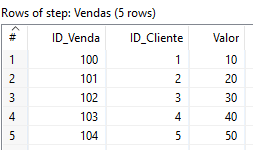


Figura - Dados do ficheiro Vendas

* **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.

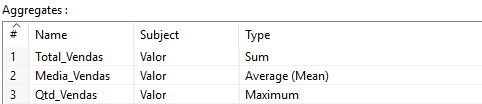


Figura - Fórmulas aplicadas

* 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.

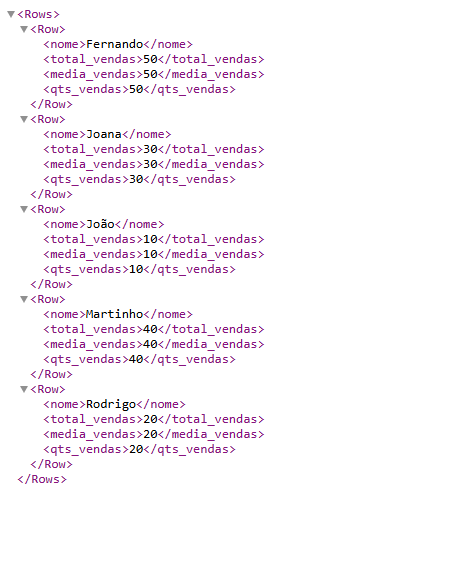


Figura - Ficheiro gerado em xml

# **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.

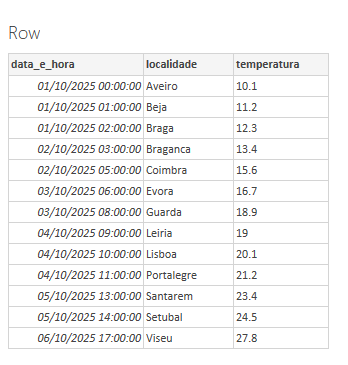


Figura - Dados retirados do xml alterado

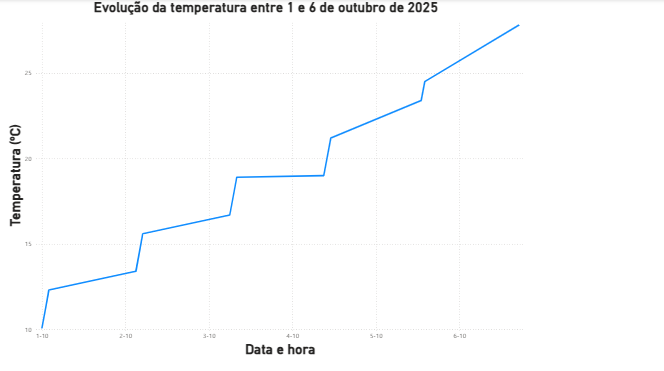


Figura - Gráfico com as temperaturas registadas

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

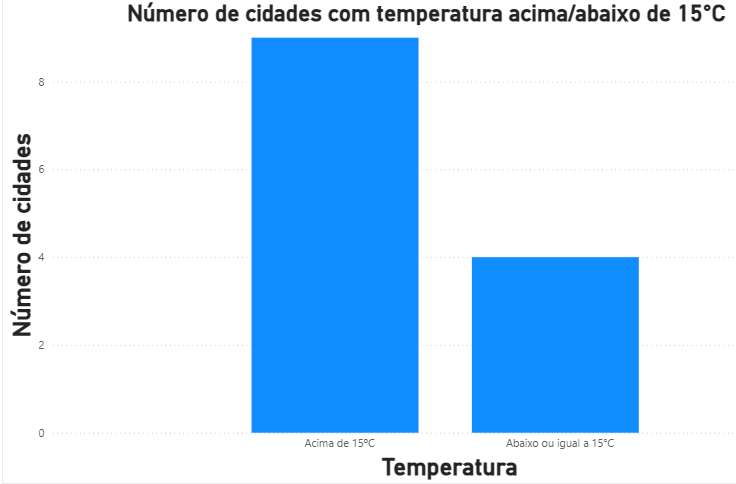


Figura - Cidades e temperaturas

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/

- <https://www.microsoft.com/pt-pt/power-platform/products/power-bi?market=pt>

- https://support.microsoft.com/pt-pt/office/xml-para-principiantes-a87d234d-4c2e-4409-9cbc-45e4eb857d44