

Análise Dádivas de Sangue

Helder Miguel da Silva Costa

Nº 29576 – Regime Pós-Laboral

Professor

Luís Ferreira

Ano letivo 2025/2026

Licenciatura em Engenharia de Sistemas Informáticos

Escola Superior de Tecnologia

Instituto Politécnico do Cávado e do Ave

RESUMO

O presente trabalho da disciplina de Integração de Sistemas de Informação (ISI) focou-se na aplicação e experimentação da ferramenta KNIME Analytics Platform em processos de ETL (Extract, Transformation and Load).

**O Processo de ETL**

O objetivo foi integrar, limpar e consolidar múltiplos fluxos de dados em formato JSON (incluindo reservas.json) sobre Reservas de Componentes Sanguíneos e Dadores (1ª Vez e Regulares).

O fluxo de trabalho no KNIME implementou as seguintes transformações principais:

* Extração de campos aninhados em JSON, incluindo coordenadas geográficas.
* Padronização de Data através de manipulação de string (join($Periodo$, "-01")) e conversão para o tipo Date&Time.
* Integração dos datasets (Reservas e Dadores) através de um nó Joiner com base nas chaves compostas Entidade e Região.
* Limpeza de dados, tratando valores omissos em colunas numéricas com o valor fixo zero.
* Carga do dataset final para um ficheiro Parquet (DadosDadiva.parquet).

**Análise e Resultados**

As visualizações obtidas a partir do dataset processado destacam a distribuição dos dados por região:

* A Soma de Reservas por Região revela um domínio claro da Entidade Central (IPST, IP) em termos de volume total.
* A análise da Soma de Dadores corrobora este padrão, mostrando que a mesma Entidade Central detém os maiores volumes de Dadores Regulares e Dadores 1ª Vez.

O projeto demonstrou o domínio das técnicas de integração de dados e o cumprimento de vários critérios de mais-valia exigidos, preparando o caminho para futuros trabalhos que explorem a orquestração de Jobs e o uso de Expressões Regulares.

ÍNDICE

[1. Introdução 5](#_Toc211411607)

[1.1. Enquadramento 5](#_Toc211411608)

[1.2. Problema 5](#_Toc211411609)

[1.3. Objetivo 5](#_Toc211411610)

[2. Problema 6](#_Toc211411611)

[2.1. Descrição do Problema 6](#_Toc211411612)

[2.2. Requisitos de Integração de Dados 6](#_Toc211411613)

[3. Estratégia Utilizada 7](#_Toc211411614)

[3.1. Processo ETL Detalhado 7](#_Toc211411615)

[4. Transformações (Diagramas e Explicação) 8](#_Toc211411616)

[4.1. Diagrama de Transformação Principal 8](#_Toc211411617)

[4.2. Detalhe da Fase de União (Joiner) 9](#_Toc211411618)

[4.3. Detalhe do Tratamento de Nulos (Missing Value) 10](#_Toc211411619)

[5. Jobs (Diagramas e Explicação) 11](#_Toc211411620)

[5.1. Diagrama do Job (Carga Final) 11](#_Toc211411621)

[6. Conclusão e Trabalhos Futuros 12](#_Toc211411622)

[6.1. Conclusão 12](#_Toc211411623)

[6.2. Trabalhos Futuros (Exploração de Novas Tecnologias) 12](#_Toc211411624)

# Introdução

## Enquadramento

Este relatório é submetido no âmbito da Unidade Curricular de Integração de Sistemas de Informação (ISI), da Licenciatura em Engenharia de Sistemas Informáticos (ESI), com o objetivo de focar a aplicação e experimentação de ferramentas em processos de ETL (Extract, Transformation and Load). O trabalho foi desenvolvido utilizando a plataforma KNIME Analytics Platform como ferramenta principal para o ETL, e recorrendo a ferramentas de visualização como o Microsoft Power BI.

## Problema

O problema abordado consiste na integração, transformação e análise de dados provenientes de múltiplas fontes, nomeadamente ficheiros JSON, que contêm informações dispersas sobre Reservas de Componentes Sanguíneos e Dadores (1ª Vez e Regulares) por diferentes Regiões e Entidades.

## Objetivo

O principal objetivo deste trabalho é desenvolver um Job de ETL robusto capaz de:

1. Extrair dados recolhidos de ficheiros JSON.
2. Transformar e Limpar os dados, padronizando campos temporais, tratando valores errados e aplicando filtros.
3. Integrar os diferentes datasets através de operações de Join baseadas nas chaves Entidade e Região.
4. Carregar o resultado final num formato estruturado (Parquet), pronto para a análise e criação de dashboards de visualização dos resultados conseguidos.

# Problema

## Descrição do Problema

O objetivo do trabalho é aplicar e experimentar ferramentas em processos de ETL (Extract, Transformation and Load), inerentes à integração de sistemas de informação ao nível dos dados.

O tema escolhido foca-se na integração, processamento e análise de dados sobre Dadores de Sangue e Reservas por Região. Este cenário enquadra-se na necessidade de análise e processamento de dados (datamining) e em contextos emergentes como smart environments (Health Care), onde a integração de soluções mais inteligentes em processos existentes é um desafio constante.

## Requisitos de Integração de Dados

O desafio principal é consolidar, limpar e harmonizar múltiplos fluxos de dados brutos, provenientes de ficheiros JSON, num único conjunto de dados estruturado, que permita análises comparativas e regionais.

# Estratégia Utilizada

O processo de ETL foi implementado na **KNIME Analytics Platform**, seguindo a metodologia:

## Processo ETL Detalhado

|  |  |  |
| --- | --- | --- |
| Fase | Ação | Nível de Detalhe e Parâmetros |
| Extração (E) | Leitura e extração de dados JSON. | O JSON Reader lê os ficheiros .json do sistema de ficheiros local. No JSON Path extraímos as os dados do mesmo, podendo colocar os mesmo por tipo (string, int…) |
| Transformação (T) | Desagrupamento de registos. | O nó Ungroup desagrupa as colunas criadas pelo JSON Path, tornando cada registo uma linha. |
| Transformação (T) | Filtragem de Linhas por Região. | O Value Row Filter inclui apenas valores nominais específicos da coluna Região. |
| Transformação (T) | Criação e Conversão do Campo Data. | O String Manipulation cria a coluna *periodotraba* concatenando o valor da coluna *Periodo* (e.g., "2015-05") com "-01", resultando em "2015-05-01" (join($Periodo$, "-01")). O nó String to Date&Time converte *periodotraba* para o tipo de dados Date&Time. |
| Integração (T) | União dos fluxos de dados. | O nó Joiner une o fluxo de dados "r*eservas*" com os fluxos de dados "dadores-de-sangue2025". A correspondência é feita por: Entidade e Região. |
| Limpeza (T) | Tratamento de valores omissos (nulos). | O nó Missing Value trata colunas numéricas (Float e Integer) preenchendo os valores em erro com um valor fixo de 0.0 e 0, respetivamente, assegurando que não há falhas em cálculos agregados. |
| Carga (L) | Escrita para o destino final. | O Parquet Writer seleciona as colunas finais e grava o dataset final no ficheiro *DadosDadiva.parquet*. |

# Transformações (Diagramas e Explicação)

## Diagrama de Transformação Principal

Uma imagem com texto, file, Tipo de letra, captura de ecrã

Os conteúdos gerados por IA podem estar incorretos.

Figura - Fluxo Knime

Explicação: O diagrama ilustra as múltiplas pipelines de ETL que extraem, transformam e, em seguida, unem os dados. É visível o processamento paralelo e a convergência antes da fase de limpeza e escrita.

## Detalhe da Fase de União (Joiner)

Uma imagem com texto, captura de ecrã, número, design

Os conteúdos gerados por IA podem estar incorretos.

Figura – Joiner

Explicação: A união dos diferentes datasets (Reservas, Dadores Regulares, etc.) é realizada através de um critério de correspondência estrito (Match: All of the following) nas colunas Entidade e Região. Isto garante que os dados de reservas são corretamente associados aos dados de dadores da mesma entidade e da mesma região geográfica.

## Detalhe do Tratamento de Nulos (Missing Value)

Uma imagem com texto, captura de ecrã, ecrã, software

Os conteúdos gerados por IA podem estar incorretos.

Figura - Missing Value

Explicação: O tratamento de valores nulos é crucial para a integridade da análise. O nó está configurado para preencher colunas de tipo Number (Float) com 0.0 e colunas de tipo Number (Integer) com 0. Isto é essencial, pois a ausência de um valor de reserva ou de dador numa determinada região/período deve ser tratada como zero e não como um valor nulo (null), permitindo assim a agregação e visualização corretas.

# Jobs (Diagramas e Explicação)

## Diagrama do Job (Carga Final)

Uma imagem com texto, captura de ecrã, software, ecrã

Os conteúdos gerados por IA podem estar incorretos.

Figura - Parquet Writer

Explicação: O fluxo de trabalho KNIME atua como um Job completo. O passo final é a escrita através do node Parquet Writer. O ficheiro de destino é C:\Users\helder costa\Desktop\ISI\DadosDadiva.parquet. O modo de escrita (overwrite) garante que o Job, quando reexecutado, substitui o ficheiro antigo, mantendo a versão mais atualizada dos dados processados.

# Conclusão e Trabalhos Futuros

## Conclusão

O projeto demonstrou a capacidade de desenvolver processos de ETL robustos, cumprindo o objetivo de consolidar, limpar e analisar múltiplos datasets de saúde. Foram cumpridos vários critérios de mais-valia, destacando-se o manuseamento de JSON e a implementação de Joins e limpeza de dados (Missing Value). A análise visual resultante valida o dataset processado, confirmando o predomínio da Entidade Central (IPST, IP) em volume de reservas e dadores.

## Trabalhos Futuros (Exploração de Novas Tecnologias)

Expressões Regulares (ER): Introduzir o uso de ER para normalização da coluna Região, tratando possíveis variações e erros de digitação, aumentando a qualidade do Merging de dados.

Orquestração e Jobs: Explorar ferramentas como Node-RED ou Apache Airflow para orquestrar e agendar a execução do Job de ETL, transformando-o num processo contínuo.

Acesso a APIs: Incorporar o acesso a APIs remotas para enriquecer os dados, por exemplo, usando as coordenadas de Lat/Long extraídas para obter informações contextuais adicionais.