Integração de Sistemas de Informação (ISI)

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 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 em JSON.
- Padronização de Data através de manipulação de string (join($Periodo$, "-01")) e conversão para o tipo Date&Time.
- Integração dos dados (Reservas e Dadores) através de um node Joiner com base nas chaves compostas Entidade e Região.
- Limpeza de dados, tratamento de valores nulos em colunas numéricas com o valor fixo zero.
- Carregamento dos dados finais para um ficheiro Parquet.

Análise e Resultados

As visualizações obtidas a partir dos dados processados 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 para futuros trabalhos que explorem a manuseamento de Jobs e o uso de Expressões Regulares.

ÍNDICE

1. Introdução

1.1. 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.

1.2. 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.

1.3. 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 aplicar filtros.
3. Integrar os diferentes dados através de operações de Join baseadas nas chaves Entidade e Região.
4. Carregar o resultado num formato estruturado (Parquet), pronto para a análise e criação de dashboards de visualização dos resultados conseguidos.

2. Problema

2.1. 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 e em contextos emergentes como smart environments (Health Care), onde a integração de soluções mais inteligentes em processos existentes é um desafio constante.

2.2. 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.

3. Estratégia Utilizada

O processo de ETL foi implementado no **KNIME Analytics Platform**, seguindo a metodologia:

3.1. 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 node 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 node String to Date&Time converte *periodotraba* para o tipo de dados Date&Time.
Integração (T)	União dos fluxos de dados.	O node 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 node 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*.

4. Transformações (Diagramas e Explicação)

4.1. Diagrama de Transformação Principal

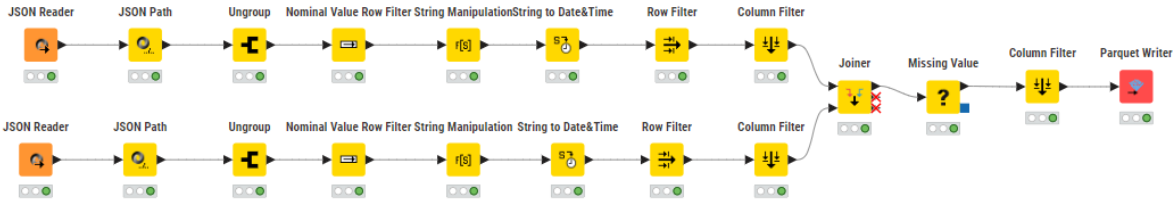

Figura 1 - Fluxo Knime

Explicação: O diagrama ilustra as múltiplas pipelines de ETL que extraem, transformam e, em seguida, unem os dados. São visíveis os processamentos paralelos e a convergência antes da fase de limpeza e escrita.

4.2. Detalhe da Fase de União (Joiner)

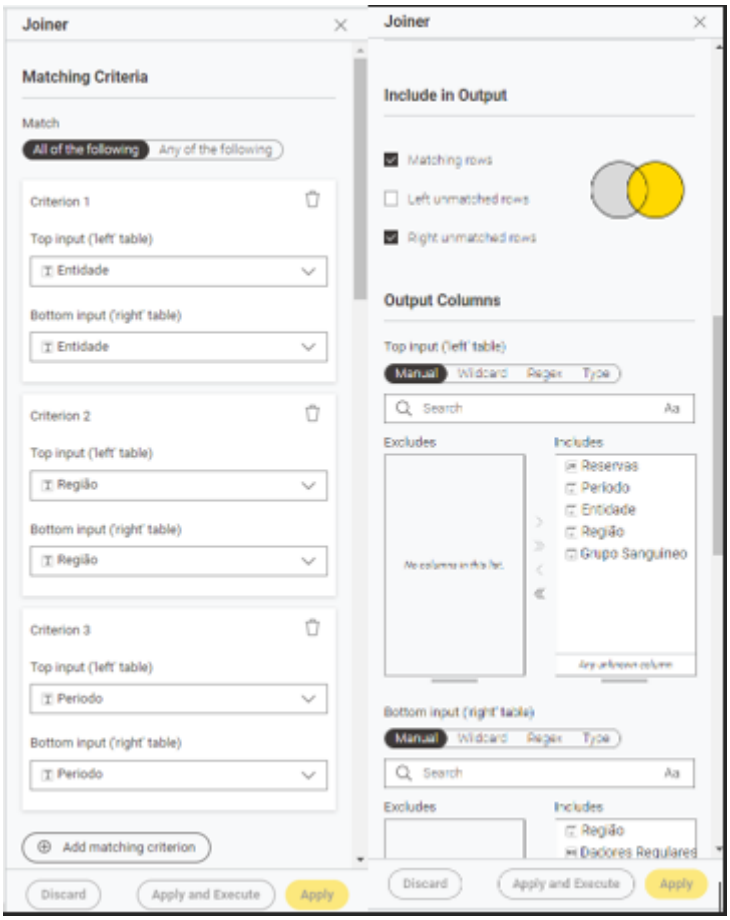

Figura 2 – Joiner

Explicação: A união dos diferentes conjuntos de dados (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.

4.3. Detalhe do Tratamento de Nulos (Missing Value)

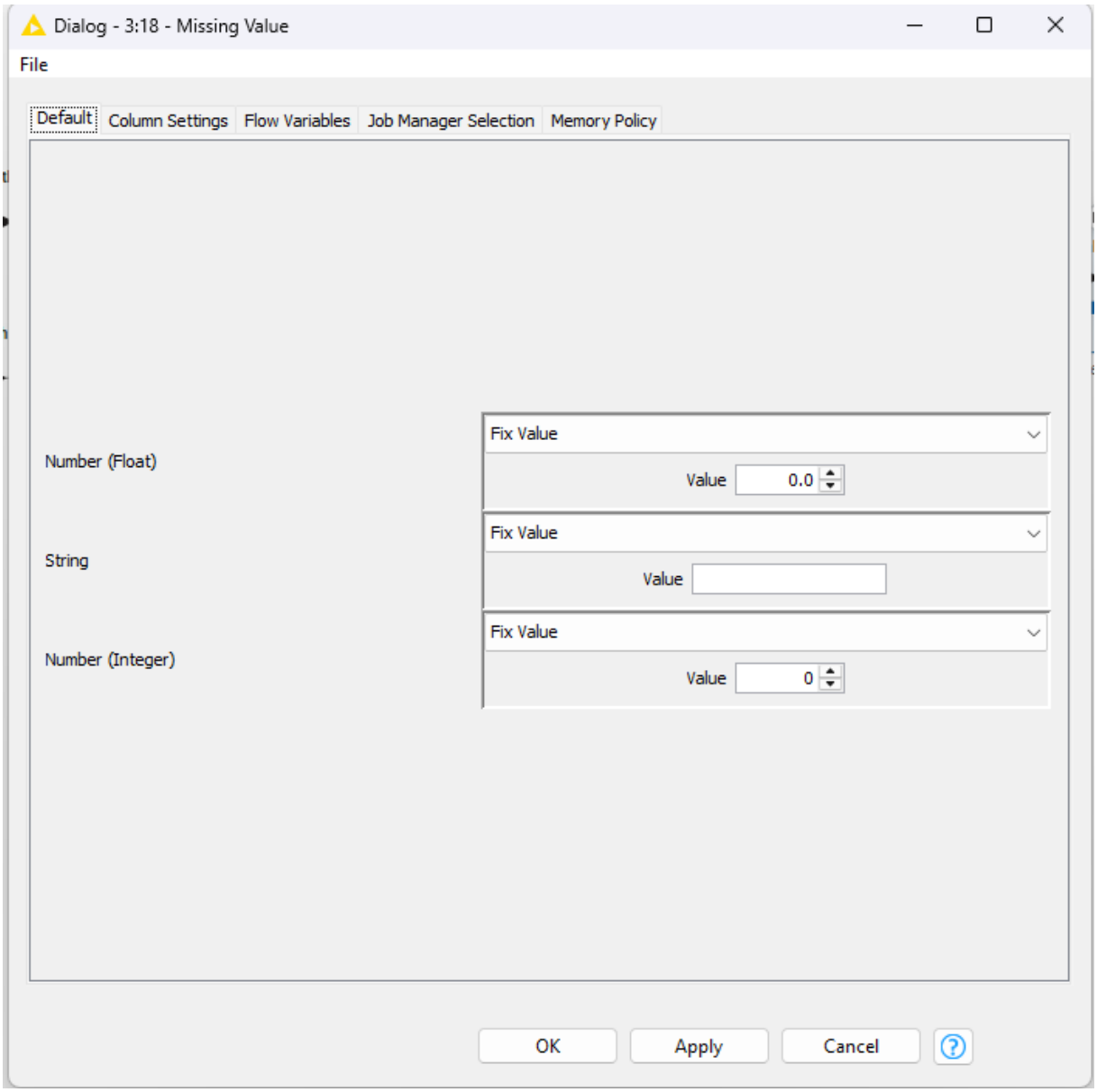

Figura 3 - Missing Value

Explicação: O tratamento de valores nulos é crucial para a integridade da análise. O node 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.

5. Jobs (Diagramas e Explicação)

5.1. Diagrama do Job (Carga Final)

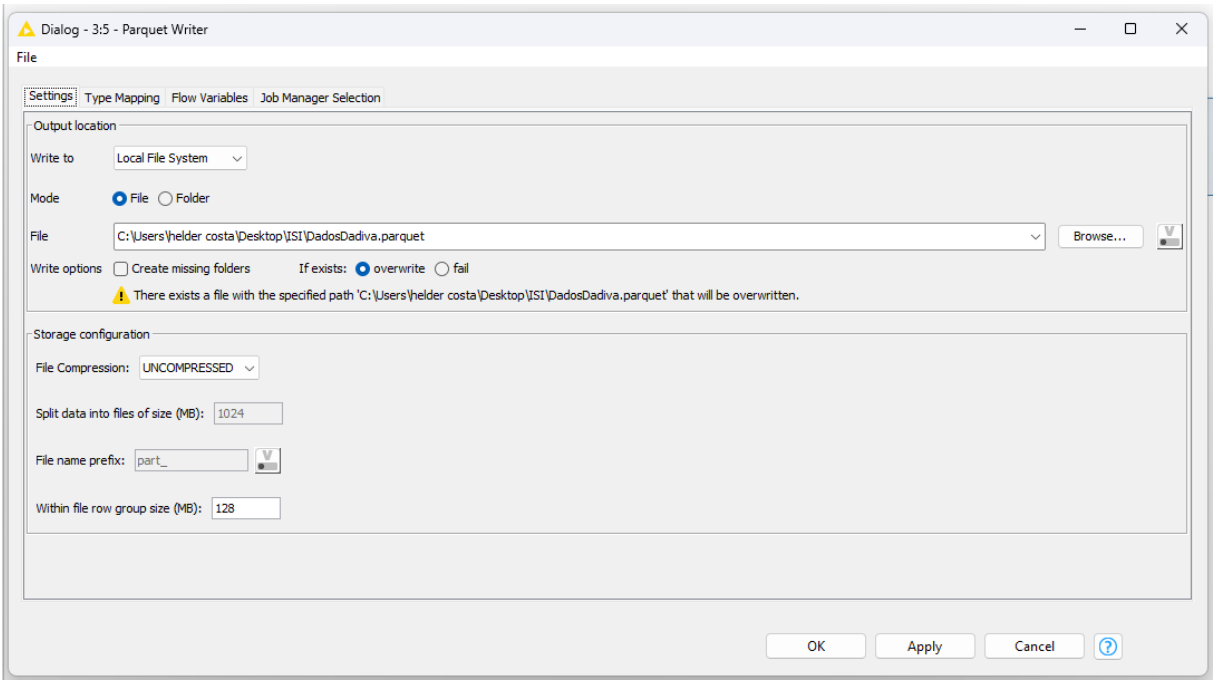

Figura 4 - Parquet Writer

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.

6. Conclusão e Trabalhos Futuros

6.1. Conclusão

O projeto demonstrou a capacidade de desenvolver processos de ETL robustos, cumprindo o objetivo de consolidar, limpar e analisar múltiplos daods de saúde. Foram cumpridos vários critérios, destacando-se o manuseamento de JSON, a implementação de Joins e limpeza de dados (Missing Value). A análise visual resultante valida os dados processados.

6.2. 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.

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.