Trabalho prático 1: ETL

Jorge Manuel Oliveira Mendes

Nº 18466 – Regime Pós-laboral

Ano letivo 2024/2025

Licenciatura em Engenharia de Sistemas Informáticos

Escola Superior de Tecnologia

Identificação do Aluno

Jorge Manuel Oliveira Mendes

Aluno nº18466, regime pós-laboral

Licenciatura em Engenharia de Sistemas Informáticos

RESUMO

Este trabalho prático teve como objetivo a implementação de um processo de ETL (Extract, Transform, Load) utilizando a plataforma KNIME, no contexto da disciplina de Integração de Sistemas de Informação. O foco foi a manipulação e transformação de um dataset contendo informações sobre cães, permitindo ao utilizador escolher uma raça específica e exportar os dados filtrados num ficheiro JSON.

O processo envolveu várias etapas, incluindo a validação de dados, filtragem e exportação. Utilizou-se o String Widget para permitir ao utilizador selecionar a raça de cão pretendida, seguido de um Row Filter para aplicar a filtragem aos dados com base na raça escolhida. Além disso, foi implementada a validação da coluna "Sexo", para garantir que apenas os valores "M" ou "F" eram aceites, e a transformação de datas de nascimento para calcular a idade dos cães.

Finalmente, os dados filtrados foram convertidos para formato JSON utilizando os nós Table to JSON e JSON Writer, concluindo o processo de exportação. Este trabalho proporcionou uma experiência prática com ferramentas de ETL e processos de manipulação de dados, contribuindo para a consolidação dos conceitos associados à integração de sistemas de informação.

ABSTRACT

This practical assignment aimed to implement an ETL (Extract, Transform, Load) process using the KNIME platform, as part of the Information Systems Integration course. The focus was on manipulating and transforming a dataset containing information about dogs, allowing the user to choose a specific breed and export the filtered data into a JSON file.

The process involved several steps, including data validation, filtering, and exporting. The String Widget was used to allow the user to select the desired dog breed, followed by a Row Filter to apply filtering based on the chosen breed. Additionally, the "Gender" column was validated to ensure only "M" or "F" values were accepted, and birthdates were transformed to calculate the dogs' ages.

Finally, the filtered data was converted to JSON format using the Table to JSON and JSON Writer nodes, completing the export process. This work provided practical experience with ETL tools and data manipulation processes, contributing to the consolidation of concepts related to information systems integration.

ÍNDICE

1. Introdução 1
1.1. Objetivos 1
1.2. Contexto 1
1.3. Estrutura do documento 2
2. Análise e Transformações 3
3. Implementação e Ferramentas 9
3.1. Ferramentas e Nós utilizados 9
3.2. Implementação do Workflow 11
3.3. Vantagens da implementação no KNIME 12
4. Resultados e Conclusão 13
4.1. Resultados 13
4.2. Conclusão 14
4.3. Trabalhos Futuros 15

ÍNDICE DE FIGURAS

Figura 1 - Extração de dados....3
Figura 2 - Validação de dados....4
Figura 3 - Transformação de datas....5
Figura 4 - Cálculo de idades....5
Figura 5 - Filtragem de Raças....6
Figura 6 - Configuração do String Widget....6
Figura 7 - Exportação dos resultados para JSON....7

Siglas e Acrónimos

CSV – Comma Separated Values (Valores Separados por Vírgula)

ETL – Extract, Transform, Load (Extrair, Transformar, Carregar)

JSON – JavaScript Object Notation (Notação de Objetos JavaScript)

KNIME – Konstanz Information Miner (Minerador de Informação de Constança)

1. Introdução

A integração de sistemas de informação tornou-se crucial para organizações que lidam com grandes volumes de dados provenientes de várias fontes. Os processos de ETL (Extract, Transform, Load) são fundamentais para garantir que os dados brutos possam ser processados, transformados e carregados em sistemas de destino para análises e tomadas de decisão. Este trabalho prático foca-se na implementação de um fluxo de ETL usando a plataforma KNIME, aplicando várias transformações e validações sobre um dataset contendo informações sobre cães.

1.1. Objetivos

O principal objetivo deste trabalho foi a criação de um fluxo de ETL completo utilizando o KNIME. Este fluxo inclui a validação de dados, filtragem de registos com base em condições específicas, cálculo de idades e exportação dos resultados em formato JSON. O utilizador pode selecionar a raça dos cães a ser analisada. O trabalho também busca consolidar o entendimento de processos de integração de dados, bem como o uso de ferramentas e tecnologias modernas de ETL.

1.2. Contexto

Este trabalho foi desenvolvido no âmbito da unidade curricular de Integração de Sistemas de Informação, no curso de Licenciatura em Engenharia de Sistemas Informáticos. O contexto do trabalho envolve a simulação de um cenário onde é necessário limpar, transformar e analisar dados de um conjunto de cães, selecionando apenas aqueles que cumprem certos critérios. O uso da plataforma KNIME permitiu a aplicação prática de conceitos teóricos, com o objetivo de facilitar a manipulação e integração de dados de forma eficiente.

1.3. Estrutura do documento

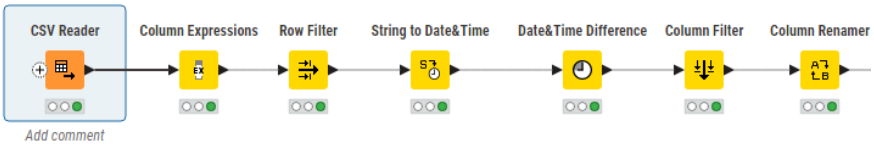
Este relatório está estruturado da seguinte forma:

- **Introdução:** Inclui o contexto e os objetivos do trabalho, bem como uma visão geral da importância dos processos de ETL.
- **Análise e Transformações:** Explica as operações realizadas nos dados, incluindo filtragem, validação, e cálculo de idades.
- **Implementação e Ferramentas:** Descreve o workflow construído no KNIME, as ferramentas utilizadas e os detalhes técnicos da implementação.
- **Resultados e Conclusão:** Apresenta os resultados obtidos com a execução do fluxo de ETL e reflete sobre as aprendizagens e possíveis melhorias futuras.

2. Análise e Transformações

Neste trabalho, foi utilizada a plataforma KNIME para realizar várias transformações no dataset de cães. O processo envolveu várias etapas de ETL (Extract, Transform, Load), com foco em validação de dados, filtragem de informações e exportação dos resultados. A seguir, são detalhadas as operações realizadas e os respectivos desafios e soluções implementadas.

1. **Extração de Dados**



ible Flow Variables

Columns: 5 Table Statistics

RowID	Nome String	Sexo String	Raca String	Pelagem String	Data_nasc String
Row0	SOPHIA	F	SRD	Marrom	07/mar/05
Row1	DRAGO	N	Pastor Belga	Preta	23/jun/97
Row2	DARA	F	Cocker Spaniel Ingles	Marrom	12/mar/95
Row3	GRETA	P	SRD	Preta	10/jun/98
Row4	MINE	F	Rottweiler	Preto e Marrom	23/mar/97

Figura 1 - Extração de dados

O dataset foi inicialmente extraído de um ficheiro CSV, que continha várias colunas com informações sobre os cães, tais como nome, data de nascimento, sexo e raça. Este ficheiro foi lido no KNIME usando o nó CSV Reader, que converteu os dados brutos numa tabela pronta para manipulação.

2. **Validação dos Dados**

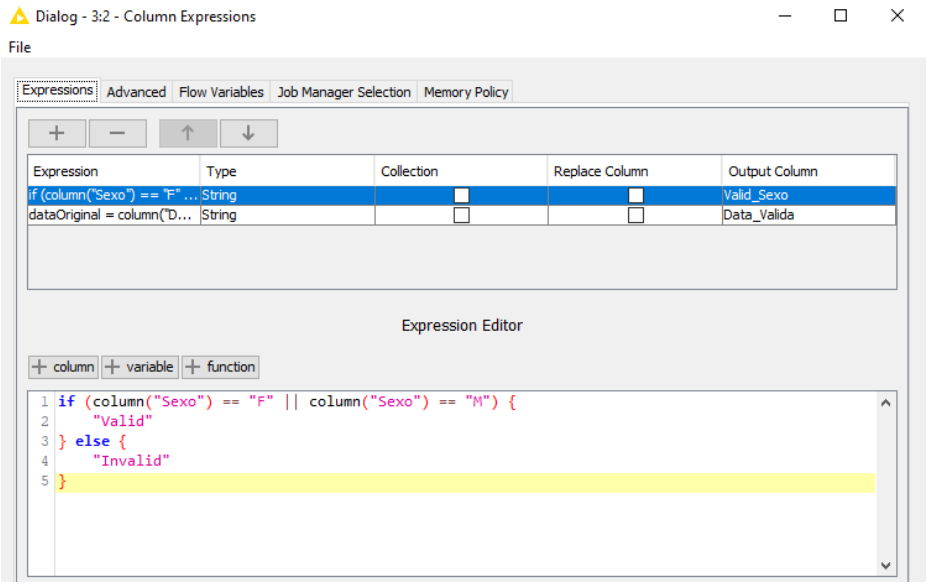


Figura 2 - Validação de dados

Foi necessário garantir que os valores na coluna Sexo fossem válidos, ou seja, apenas "M" (masculino) ou "F" (feminino). Para isso, utilizou-se o nó Column Expressions para verificar cada valor e substituir os inválidos. Caso o valor fosse diferente de "M" ou "F", a linha correspondente foi marcada como inválida e removida posteriormente do dataset, assegurando a consistência dos dados.

3. Transformação de Datas e Cálculo de Idade

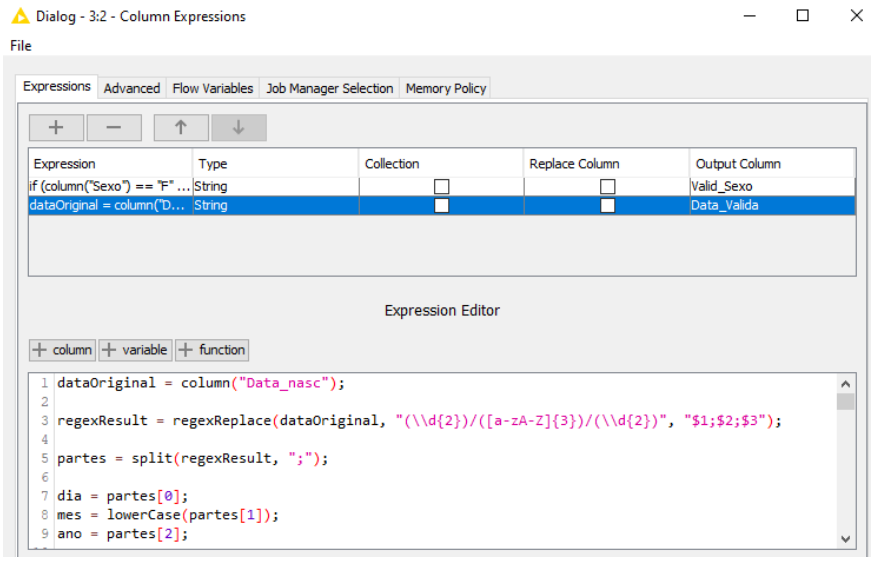


Figura 3 - Transformação de datas

A coluna Data de Nascimento apresentava as datas no formato dd/mmm/yy, onde os meses eram abreviados (ex: "jan", "mar"). Foi necessário converter este formato para dd/MM/yyyy utilizando o nó Column Expressions e uma lógica de expressões regulares para substituir os meses abreviados por números correspondentes.

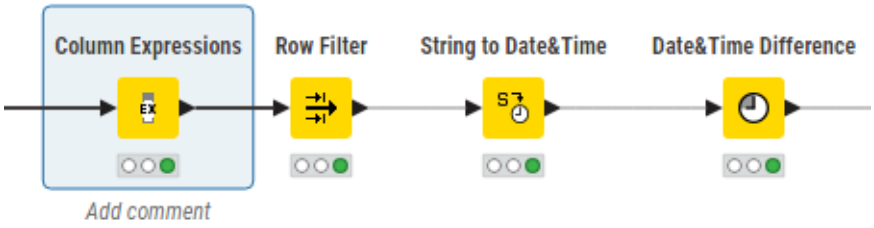


Figura 4 - Cálculo de idades

Em seguida, foi calculada a idade dos cães com base na data de nascimento, comparando-a com a data atual. O nó Date&Time Difference foi utilizado para calcular a diferença em anos entre as datas.

4. **Filtragem de Raças**

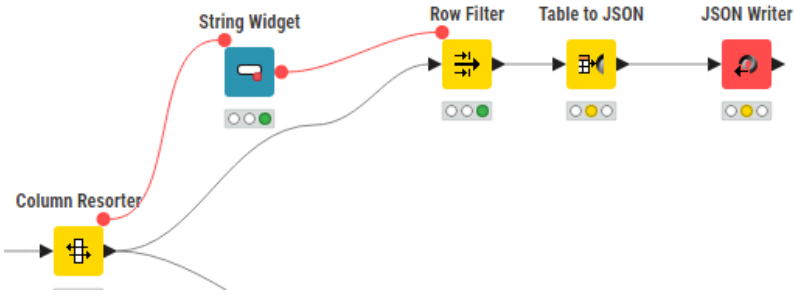


Figura 5 - Filtragem de Raças

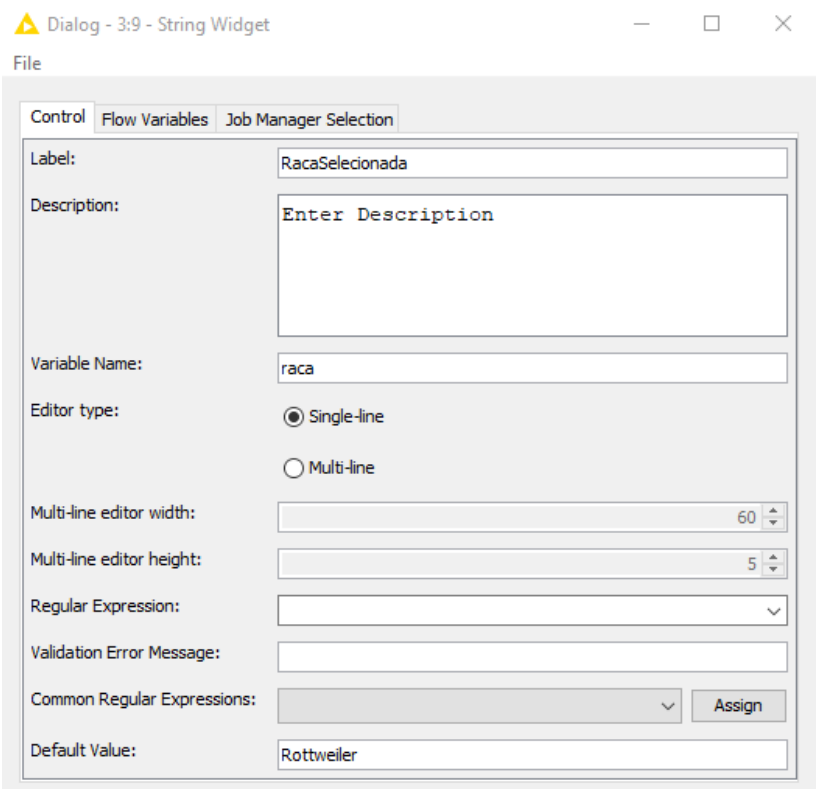


Figura 6 - Configuração do String Widget

Para tornar o processo interativo, foi utilizado o nó String Widget, que permitiu ao utilizador selecionar uma raça específica de cães para análise. A partir dessa seleção, o nó Row Filter foi aplicado para filtrar o dataset e manter apenas os cães da raça escolhida. Esta etapa tornou o fluxo flexível, permitindo que o utilizador modifique o critério de seleção conforme necessário.

5. **Exportação dos Resultados para JSON**

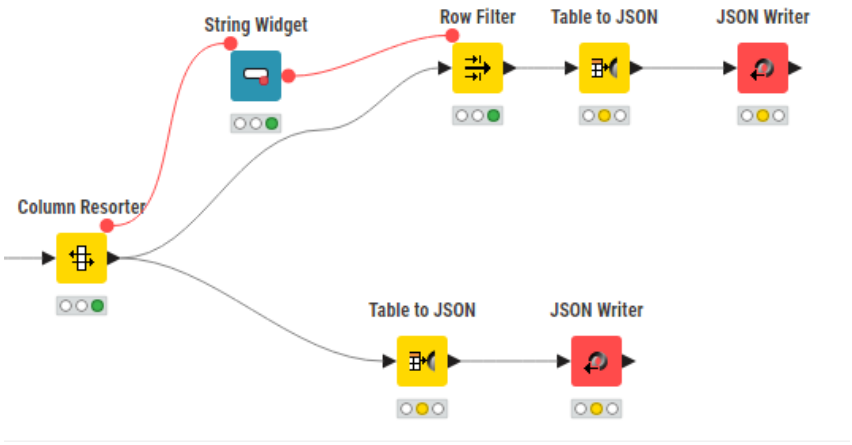


Figura 7 - Exportação dos resultados para JSON

Após a filtragem e transformação dos dados, os resultados foram exportados em formato JSON utilizando os nós Table to JSON e JSON Writer. Isso permitiu que os dados fossem guardados e transferidos de forma eficiente, mantendo um formato leve e amplamente utilizado em aplicações de sistemas de informação.

3. Implementação e Ferramentas

O desenvolvimento deste projeto foi realizado na plataforma KNIME, uma ferramenta poderosa e versátil para a criação de fluxos de trabalho em processos de ETL (Extract, Transform, Load). O KNIME permite a construção visual de workflows, integrando uma variedade de nós que facilitam a manipulação, transformação e exportação de dados. Abaixo, são detalhados os principais nós e ferramentas utilizadas no fluxo de trabalho para atender aos requisitos do projeto.

3.1. Ferramentas e Nós utilizados

Abaixo estão as ferramentas e principais nós usados no fluxo de trabalho do KNIME:

1. **CSV Reader:**
   - Utilizado para carregar o dataset inicial em formato CSV para o KNIME.
   - Este nó lê os dados originais, transformando-os numa tabela KNIME para processamento subsequente.

2. **String Widget:**
   - Este nó permitiu a criação de uma interface interativa para o utilizador escolher uma raça de cão específica para filtragem.
   - A raça selecionada pelo utilizador foi armazenada como uma variável de fluxo, usada depois no filtro de dados.

3. **Row Filter:**
    - Implementado em dois momentos diferentes do fluxo:
        - Primeiro, para filtrar os registos com que tinham um género válido.
        - Depois, para filtrar os registos com base na raça selecionada pelo utilizador através do String Widget.
    - Este nó permitiu um controlo eficiente das informações processadas, assegurando que apenas os registos relevantes fossem mantidos.

4. **Column Expressions:**
- Utilizado para validação e transformação de dados, especialmente para a coluna "Data de Nascimento" e "Sexo".
- Realizou a validação dos valores na coluna Sexo, assegurando que apenas "M" e "F" fossem valores aceitáveis.
- Também foi utilizado para converter o formato das datas de nascimento de dd/mmm/yy para dd/MM/yyyy, permitindo um cálculo correto da idade.

5. **Date&Time Difference:**
- Aplicado para calcular a idade dos cães, comparando a coluna "Data de Nascimento" com a data atual.
- Este nó produziu uma coluna adicional que contém a idade dos cães.

6. **Table to JSON e JSON Writer:**
- Estes dois nós foram usados para exportar os dados filtrados no formato JSON, conforme o requisito do projeto.
- Table to JSON converteu a tabela filtrada numa estrutura JSON, enquanto o JSON Writer guardou os dados num ficheiro JSON final, permitindo a sua fácil utilização em outros sistemas.

3.2. Implementação do Workflow

O workflow de ETL foi implementado em etapas, cada uma representando um processo específico de manipulação de dados:

1. **Carregamento e Preparação dos Dados:**
    - O ficheiro CSV inicial foi carregado para o KNIME e transformado numa tabela KNIME para facilitar a manipulação dos dados. Esta tabela serviu como a base para todas as operações subsequentes.

2. **Interatividade com o Utilizador:**
    - O nó String Widget foi integrado para permitir a seleção interativa da raça de cães a ser filtrada. Esta escolha foi posteriormente aplicada através de um Row Filter para reter apenas os registos que correspondiam à raça escolhida.

3. **Transformação e Validação:**
    - A coluna "Sexo" foi validada para garantir que apenas valores "M" ou "F" fossem incluídos, enquanto as datas de nascimento foram convertidas para um formato adequado ao cálculo da idade.
    - A idade foi calculada usando a diferença entre a data de nascimento e a data atual, gerando uma coluna de idade em anos para cada cão.

4. **Exportação:**
    - Finalmente, os dados foram convertidos para JSON usando os nós Table to JSON e JSON Writer, e exportados para um ficheiro JSON, permitindo a fácil integração dos dados em outras aplicações.

3.3. Vantagens da implementação no KNIME

O uso do KNIME para este projeto ofereceu várias vantagens:

- Interface visual e intuitiva: facilitou a criação e modificação dos workflows.
- Flexibilidade na manipulação de dados: através dos diversos nós disponíveis para transformação, filtragem e validação.
- Exportação para múltiplos formatos: a capacidade de exportar dados para JSON é especialmente útil em ambientes que exigem interoperabilidade entre sistemas.

4. Resultados e Conclusão

4.1. Resultados

A implementação do processo de ETL no KNIME produziu resultados eficazes e alinhados com os objetivos propostos. Os principais resultados alcançados foram:

1. **Interatividade na Escolha de Dados:**
    - O uso do String Widget permitiu ao utilizador selecionar interativamente a raça de cães para análise, proporcionando um nível de flexibilidade importante. Este componente adicionou valor ao processo ao permitir que o utilizador filtrasse dinamicamente os dados conforme a necessidade.

2. **Transformação e Limpeza de Dados:**
    - As validações aplicadas aos dados, como a verificação dos valores na coluna Sexo e a conversão do formato das datas, garantiram a integridade e consistência dos dados. A transformação da data de nascimento para o formato dd/MM/yyyy e o cálculo da idade foram realizados com sucesso.
    - A limpeza dos dados através do Row Filter assegurou que o dataset final continha apenas os registos relevantes, conforme os critérios estabelecidos.

3. **Exportação para JSON:**
    - O processo de exportação produziu um ficheiro JSON contendo exclusivamente os registos dos cães que atendiam aos critérios definidos (raça selecionada pelo utilizador). Esta exportação em JSON facilita a integração dos dados em outros sistemas, tornando o workflow útil para ambientes que requerem formatos interoperáveis.

4. **Facilidade na Manutenção e Extensibilidade do Workflow:**
   - A modularidade e clareza do workflow desenvolvido no KNIME permitem que o processo seja facilmente mantido e expandido. Novos critérios de filtragem, formatos de exportação ou operações de transformação podem ser adicionados sem necessidade de reconstruir o fluxo do zero.

4.2. Conclusão

A execução deste projeto no KNIME demonstrou a eficácia e flexibilidade da plataforma para processos de ETL. A criação de um fluxo de trabalho visual permitiu implementar transformações complexas de dados de forma intuitiva e eficiente. As principais aprendizagens incluem:

- **Experiência prática em ETL:** O projeto consolidou os conhecimentos teóricos sobre processos de ETL, desde a extração e transformação até ao carregamento dos dados.
- **Uso de ferramentas de integração de dados:** KNIME mostrou-se uma ferramenta poderosa para manipulação e integração de dados, com uma variedade de nós que suportam operações avançadas de filtragem, validação e exportação.
- **A importância da validação e limpeza de dados:** Através da validação do conteúdo das colunas e da eliminação de registos fora dos critérios, foi possível manter um dataset preciso e adequado ao objetivo final.

4.3. Trabalhos Futuros

Para evoluir o projeto, identificaram-se algumas melhorias e expansões possíveis:

- **Integração com APIs externas:** Poderia ser interessante expandir o workflow para incorporar dados de fontes externas em tempo real, como dados de saúde animal ou demográficos sobre raças.
- **Automatização de Relatórios:** Adicionar uma camada de visualização ou geração de relatórios automáticos no KNIME, com gráficos ou tabelas sobre os resultados do filtro de dados.
- **Suporte a Múltiplos Formatos de Exportação:** Implementar a exportação em outros formatos (por exemplo, XML ou Excel), aumentando a versatilidade do workflow para diferentes necessidades de integração de dados.

Este projeto constituiu uma experiência prática valiosa em processos de ETL e manipulação de dados, fortalecendo a compreensão sobre os desafios e soluções associados à integração de sistemas de informação.

Bibliografia

KNIME Documentation. (n.d.). KNIME Analytics Platform User Guide. KNIME. Disponível em: https://docs.knime.com

KNIME Extensions and Integrations. (n.d.). KNIME Extensions Guide. KNIME Hub. Disponível em: https://hub.knime.com