Relatório do trabalho da cadeira de Integração de Sistemas de Informação

Projeto ETL

Carlos Moreda – nº 26875 - a26875@alunos.ipca.pt

GitHub - Carlos26875/ISI

ESI-PL

Outubro de 2024

Afirmo por minha honra que não recebi qualquer apoio não autorizado na realização deste trabalho prático. Afirmo igualmente que não copiei qualquer material de livro, artigo, documento web ou de qualquer outra fonte exceto onde a origem estiver expressamente citada.

Carlos Moreda – nº 26875

Conteúdo

Introdução

O presente trabalho enquadra-se na unidade curricular de Integração de Sistemas de Informação (ISI) e tem como objetivo explorar os processos ETL, essenciais para a gestão eficiente de dados. No atual cenário em que a quantidade de dados disponíveis cresce exponencialmente, torna-se crucial desenvolver competências que permitam extrair informações relevantes a partir de múltiplas fontes de dados.

Este projeto foca-se na análise automóvel, um tema de grande importância, uma vez que a indústria automotiva enfrenta desafios constantes relacionados à competitividade, sustentabilidade e inovação tecnológica. Através da análise de dados de veículos, é possível obter insights valiosos sobre tendências de consumo, eficiência energética e preferências dos consumidores, contribuindo para as tomadas de decisões tanto por parte dos consumidores quanto dos gestores.

Para este trabalho, utilizei a ferramenta KNIME, que permite realizar operações complexas de manipulação e análise de dados de forma intuitiva e visual. A metodologia adotada envolve a leitura, limpeza e junção de dados do mercado automóvel. Através da implementação dos processos ETL, busquei não apenas garantir que os dados sejam precisos e relevantes, mas também facilitar a tomada de decisões sobre a aquisição e utilização de veículos.

Problema

Estratégia

Para abordar o problema de integração e análise de dados de veículos, a estratégia adotada no projeto inclui várias etapas, cada uma das quais crucial para o sucesso da análise e para a obtenção de resultados significativos. A seguir, detalho as fases da estratégia, que se alinha aos processos ETL:

1. **Leitura de Dados**:
    - **Fontes de Dados**: Utilização de diferentes fontes de dados, incluindo ficheiros CSV e API JSON, para garantir uma variedade de informações sobre veículos.
    - **Leitura de Dados**: Implementação de um processo de extração que permita coletar dados relevantes sobre preços, consumo, marcas e modelos de veículos, garantindo que as informações sejam atualizadas e abrangentes.
2. **Transformação de Dados**:
    - **Limpeza de Dados**: Aplicação de técnicas de normalização e validação para tratar valores ausentes e formatar dados de forma consistente. Por exemplo, utilização de expressões regulares para categorizar tipos de combustível e criar colunas com informações processadas.
    - **Integração de Dados**: Junção dos dados provenientes de diferentes fontes, utilizando operações de "Joiner" e "Group by" no KNIME, a fim de criar um conjunto de dados unificado e coerente.
    - **Cálculos e Análises**: Desenvolvimento de fórmulas e regras para calcular métricas relevantes, como custo por quilômetro e categorização de veículos em "Econômico", "Médio" e "Caro". Essas análises fornecerão insights valiosos sobre o mercado automotivo.
3. **Geração de Logs**:
    - **Registo de Processos**: Implementação de um sistema de logs para documentar as operações realizadas no fluxo de trabalho, facilitando a rastreabilidade e a auditoria dos dados. Os logs incluirão informações sobre veículos processados, categorias atribuídas.
4. **Armazenamento de Dados**:
    - **Armazenamento de Dados**: Após o processamento e a transformação, os dados serão exportados para diferentes formatos, como XML e CSV, permitindo sua utilização em outras aplicações ou para relatórios.

Desenvolvimento

Transformações

Figura 1 - Leitura de dados

Esta fase é a base do processo, responsável por reunir as fontes de dados. O **Get Request** e **JSON Path** são usados para obter dados da API JSON, enquanto o **CSV Reader** importa dados locais de um ficheiro.csv. Ambas as fontes são combinadas com o nó **Joiner** para integrar as informações em uma única tabela, permitindo o processamento conjunto.

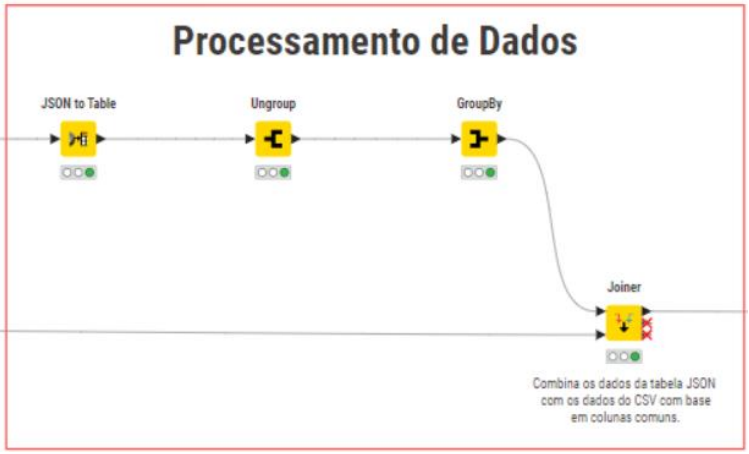

Figura 2 - Processamento de dados

Aqui ocorre a transformação dos dados estruturados em JSON para tabela, utilizando o nó **Json to Table**. O nó **Ungroup** permite separar dados, simplificando a visualização. Em seguida, o nó **Group by** agrupa os dados com base em atributos como marca, e o **Joiner** integra novamente os dados transformados e agrupados.

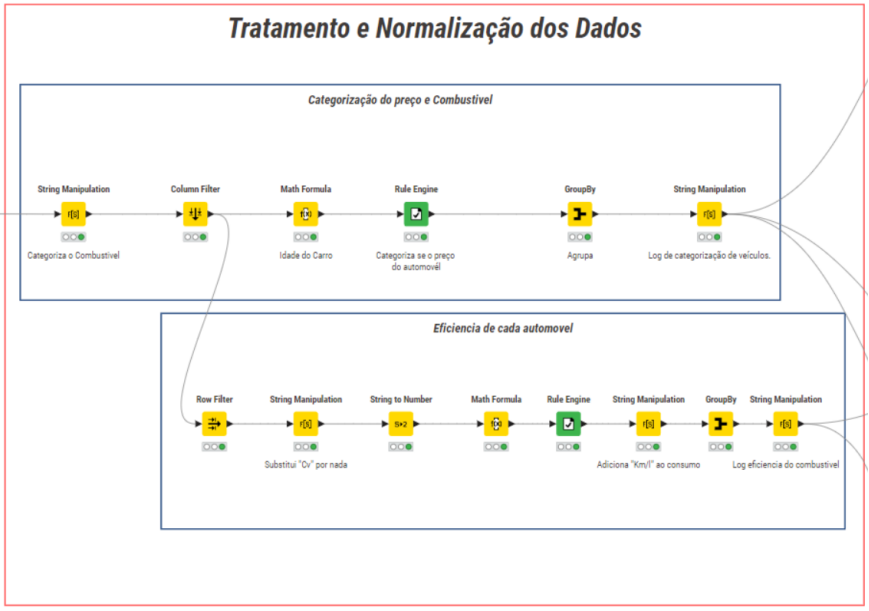

Figura 3 - Tratamento e Normalização dos dados

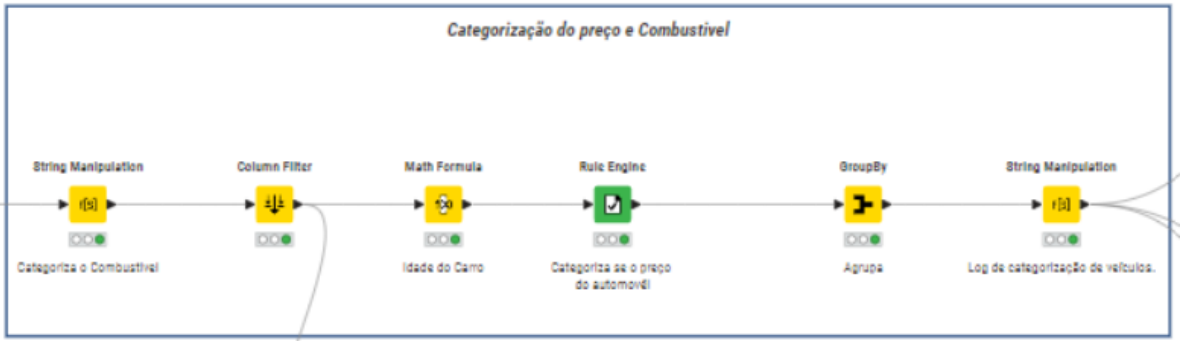

Figura 4 - Transformações de Categorização do preço e combustível

Nesta fase, os dados passam por um processo de categorização e normalização. Inicialmente, o nó **String Manipulation** é utilizado para categorizar os tipos de combustível dos veículos, facilitando a análise. Após essa etapa, o **Column Filter** é aplicado para selecionar apenas as colunas relevantes para a análise seguinte. Em seguida, o **Rule Engine** classifica os preços dos automóveis em três categorias: "Barato" (preço < 20.000), "Econômico" (20.000 <= preço < 30.000) e "Caro" (preço >= 30.000). O nó **Group by** é usado para agregar os dados conforme necessário, e, por fim, uma nova coluna de log é criada com o **String Manipulation**, permitindo registar as categorizações realizadas para cada veículo.

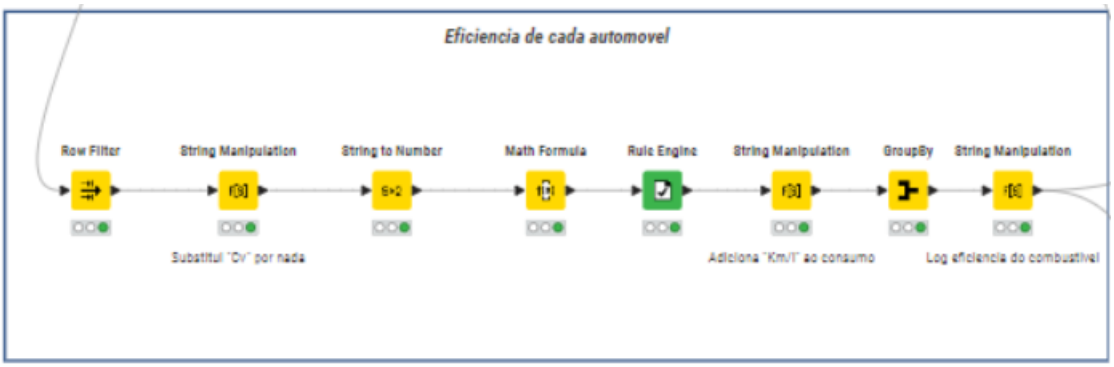

Figura 5 - Transformações Eficienia Automovel

Após a categorização de preço, é realizada a análise da eficiência dos automóveis. O **Column Filter** é usado novamente para selecionar as colunas essenciais para esta avaliação. O nó **String Manipulation** substitui o "cv" na coluna de potência por uma string vazia, limpando os dados. Em seguida, a coluna de potência é convertida para um número com o **String to Number**. O **Math Formula** é aplicado para calcular o custo por quilômetro, utilizando a fórmula que divide o preço pelo consumo. A coluna de consumo é então tratada com outro **String Manipulation**, que adiciona a unidade "Km/l" ao valor do consumo. O nó **Group by** agrega os dados por atributos relevantes, permitindo visualizar a eficiência média dos automóveis. Finalmente, é gerado um log de eficiência do combustível usando novamente o **String Manipulation**, registando a eficiência de cada veículo processado.

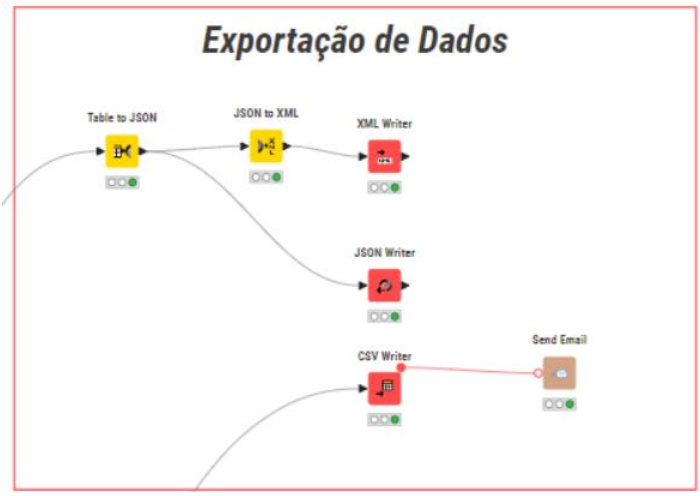

Figura 6 - Exportação de Dados

Após os logs de eficiência serem gerados, eles são exportados em formato CSV com o **CSV Writer**, e esses dados são enviados por email usando o nó de **Send Email**. Em paralelo, os logs de categorização de veículos são convertidos para **JSON** e depois para **XML** usando os nós **Table to Json** e **Json to XML**, respectivamente, antes de serem armazenados em um ficheiro XML com o **XML Writer**.

Jobs

Um "Job" no contexto deste trabalho refere-se a um conjunto de operações automatizadas realizadas sobre os dados para atingir os objetivos de processamento, integração e análise, deixo abaixo os 2 jobs mais importantes presentes no trabalho:

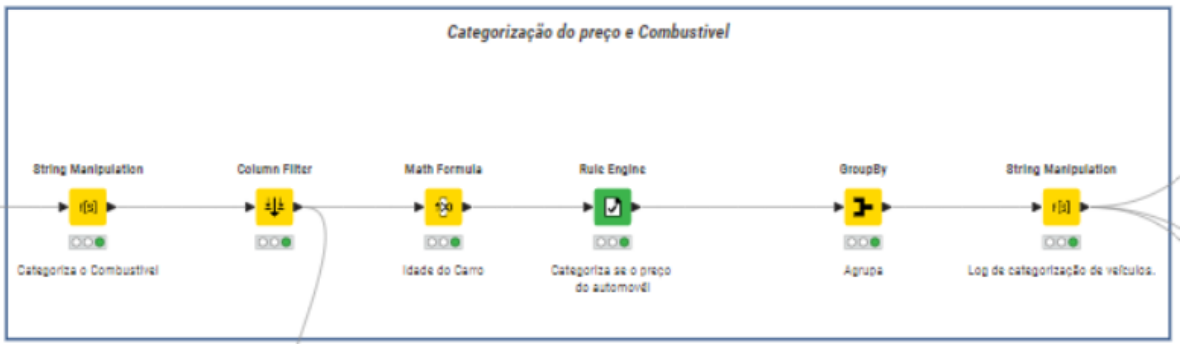

Figura 7 - Job Categorização do preço e combustível

Este job visa classificar os veículos com base no tipo de combustível e no preço, permitindo uma análise clara das opções disponíveis no mercado.

Leitura de Dados:

- Os dados são importados de fontes como CSV e API, integrando informações relevantes sobre cada veículo, incluindo marca, modelo, tipo de combustível e preço.

Tratamento de Dados

- **String Manipulation** Cria uma coluna onde categoriza o tipo de combustível dos veículos a diesel e gasolina para combustão.

- **Column Filter**: Filtra as colunas relevantes, removendo aquelas que não são necessárias para a análise, simplificando o conjunto de dados

Cálculo de Idade:

- **Math Formula:** Calcula a idade do veículo com base no ano de fabricação e no ano atual.

Categorização de Preço:

- **Rule Engine** Classifica os veículos em três categorias de preço com base no valor do veículo Economico, Médio e Caro

Agrupamento de Dados:

- **Group By**: Agrupa os dados conforme a categoria de combustível e a faixa de preço, permitindo que os dados sejam mais precisos.

Geração de Logs:

- **String Manipulation:** Cria logs que documentam a categorização dos veículos, incluindo informações como Marca, Modelo, Preço, Categoria de Combustível, Categoria de Preço

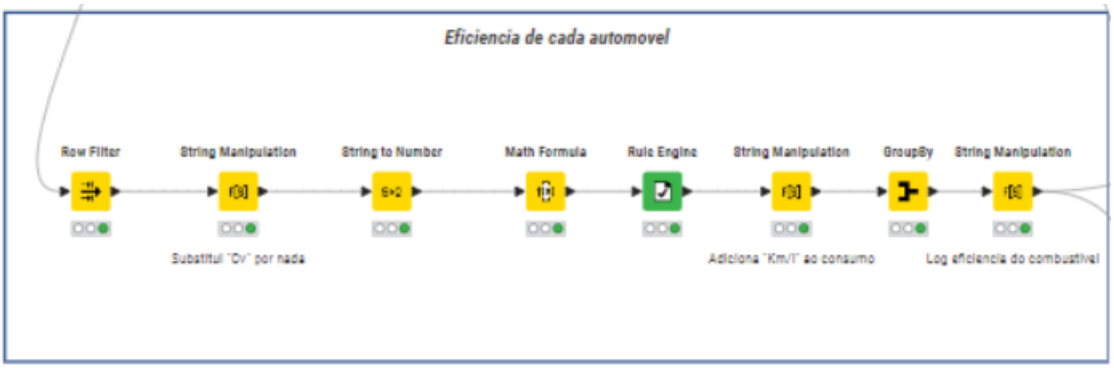

Figura 8 – Job Eficiência Automóvel

Este job visa calcular e registar a eficiência de cada veículo a combustão em termos de consumo de combustível, permitindo uma análise detalhada de desempenho.

Leitura de Dados:

- **Row Filter:** Seleciona apenas os veículos a combustão, excluindo os elétricos, garantindo que a análise se concentre apenas nas opções relevantes.

Tratamento de Dados:

- **String Manipulation:** Remove a unidade "cv" da coluna de potência, limpando os dados e facilitando a manipulação posterior.
- **String to Number:** Converte a coluna de potência para o formato numérico, permitindo que os cálculos sejam realizados sem problemas.

Cálculo de Eficiência:

- **Math Formula:** Calcula a eficiência de cada veículo utilizando a fórmula: potência / consumo.
- **Rule Engine:** Classifica a eficiência em categorias como "Muito Eficiente", "Eficiente" e "Pouco Eficiente", com base em critérios predefinidos.
- **String Manipulation:** Formata a coluna de consumo para incluir a unidade "Km/l", tornando os dados mais compreensíveis e apresentáveis.

Geração de Logs:

- **String Manipulation:** Cria logs que documentam a eficiência do combustível para cada veículo, compilando informações como Marca, Modelo, Potencia, Consumo, Categoria eficiência

Conclusão

Findo o trabalho, este me permitiu perceber a importância dos processos ETL na gestão eficiente de dados, essencial para a extração, transformação e carregamento de dados. O processo de Extração garantiu que dados relevantes sejam obtidos, enquanto a Transformação foi crucial para tratar e normalizar os dados, permitindo uma análise mais precisa e significativa.

Além disso, a definição de Jobs para automatizar tarefas torna o fluxo de trabalho mais eficiente, facilitando a categorização, cálculos e a geração de logs, que documentam o processo garantiram a rastreabilidade das operações. O carregamento dos dados, por sua vez, assegurou que os dados fiquem prontos e acessíveis para utilização em relatórios e análises.

Em projetos futuros, a implementação eficaz de processos ETL será fundamental para garantir a integridade e a qualidade dos dados.

Bibliografia

Extract, transform, load. (s.d.). Obtido de https://pt.wikipedia.org/wiki/Extract,_transform,_load

knime. (s.d.). Obtido de knime: https://www.knime.com/