Relatório Integração Sistemas

Pentahoo Data Integration

David Martinho nº25620

Instituto Politécnico do Cávado e do Ave

2024/2025

Licenciatura em Engenharia Sistemas Informáticos

3ºano

Indice de Figuras

Indice

Enquadramento

Neste trabalho desenvolvido no âmbito da disciplina de Integração de Sistemas de Informação (ISI), explora-se a aplicação de processos de *ETL* (Extract, Transform, Load) como uma solução para desafios comuns de integração e enriquecimento de dados.

Através da ferramenta Pentaho Kettle (*PDI*), foram desenvolvidos processos para transformar dados brutos, combiná-los a partir de diversas fontes, e exportá-los em formatos interoperáveis, como JSON e XML.

Além das operações de transformação e normalização, o projeto inclui a integração com serviços externos para enriquecer os dados e funcionalidades automatizadas, como o envio de relatórios via e-mail, e o uso de serviços remotos (*API*).

Estas operações refletem os cenários reais de integração de sistemas, onde a capacidade de consolidar e de enriquecer dados é essencial para responder às várias necessidades de análise e processamento em ambientes empresariais dinâmicos.

Problema

No contexto da integração de dados nas indústrias de entretenimento, as plataformas como a Netflix enfrentam o desafio de consolidar, transformar e enriquecer grandes volumes de dados sobre filmes e séries para análises, recomendações e tomada de decisão.

Este trabalho tem como objetivo demonstrar um processo de ETL específico para dados de filmes e séries, utilizando um conjunto de dados da Netflix.

O projeto envolve a extração dos dados de filmes e séries a partir de um arquivo CSV, seguido da transformação e enriquecimento das informações.

A integração desses dados enriquecidos visa melhorar a análise e a consistência das informações, permitindo uma visão mais completa e atualizada do catálogo.

Além disso, este processo inclui também etapas de limpeza e normalização dos dados, remoção de caracteres especiais e formatação de campos, além da exportação dos dados transformados para um formato estruturado em XML e a criação de um sistema de notificação via e-mail para compartilhar o relatório final.

Este fluxo de trabalho automatizado demonstra um exemplo prático de como processos de ETL podem ser aplicados para resolver desafios de integração de dados nos setores de streaming e entretenimento.

Estratégia utilizada

Leitura e Importação de Dados

Utilizou-se o operador de importação CSV para carregar os dados originais dos filmes e séries. Esse ficheiro continha informações básicas como título, data de lançamento, duração, género, descrição, entre outros atributos.

Limpeza e Normalização dos Dados

Foram aplicadas expressões regulares (ER) para remover os caracteres especiais e padronizar campos, como títulos e descrições. Também foram utilizados operadores para ajustar as datas e a duração dos filmes, criando uma estrutura uniforme.

Enriquecimento dos Dados com API Externa

Integrámos uma *API* pública para obter informações adicionais, como classificações atualizadas. Essa integração envolveu o uso de operadores REST para obter respostas JSON, que foram depois transformadas e combinadas com os dados originais.

Transformações e Joins

Para combinar os dados do CSV com as informações da API, realizámos joins baseados no título dos filmes. Assim, os novos dados ficaram incorporados. Os operadores *Append Stream* e *Join Rows* foram essenciais para esta etapa.

Exportação e Serialização para XML

Após o processamento, exportámos o resultado para XML para possibilitar futuras integrações e facilitar a visualização. Utilizamos o operador de output XML para serializar o arquivo final.

Automação de Notificações por E-mail

Para o envio automático dos dados transformados, configurámos o envio por e-mail. O operador de envio de e-mail incorporou o arquivo transformado, automatizando a entrega do resultado.

Transformações

Netflix_Movies

Esta transformação de dados foi projetada para tratar e normalizar um conjunto de dados de filmes e séries da Netflix, garantindo uma estrutura mais organizada e padronizada para análises futuras.

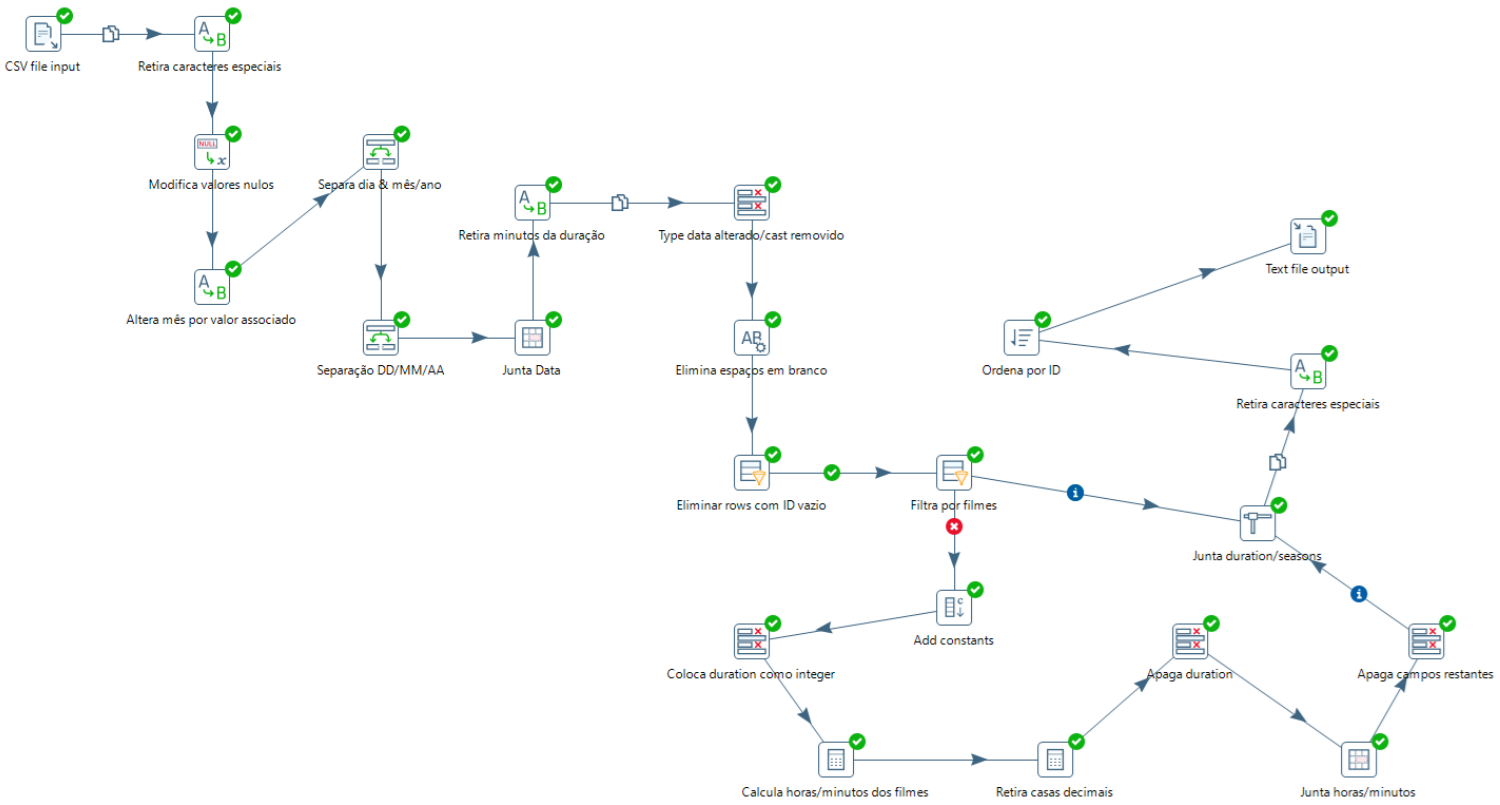

Figura 1 - Netflix_Movies.ktr

O objetivo é aplicar um conjunto de operações ETL para limpar, transformar e formatar os dados, de forma que as informações essenciais, como duração, data de lançamento e identificadores, estejam em conformidade com as necessidades do projeto.

Principais Etapas da Transformação:

Limpeza de Dados:

Remoção de caracteres especiais e espaços em branco.

Padronização dos campos nulos e tratamento de valores ausentes, ao preencher determinados campos importantes para evitar inconsistências.

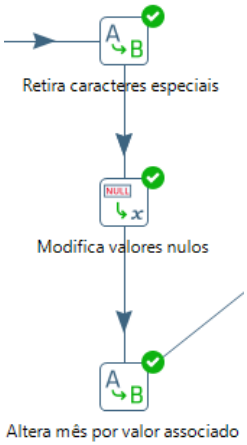

Figura 2 - Limpeza de Dados

Transformação de Campos de Data e Duração:

Separação e formatação de campos de data (dia, mês, ano) e ajuste para o formato desejado.

Conversão da duração para valores numéricos (horas e minutos), permitindo cálculos e análises mais precisas.

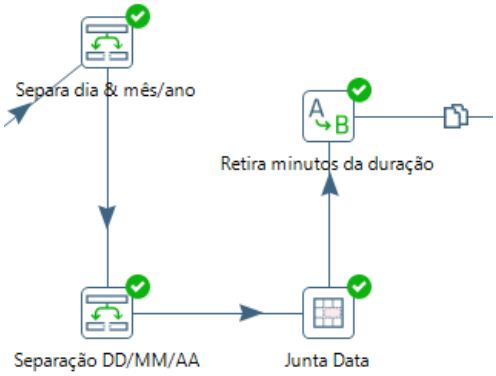

Figura 3 - Transformação de Campos

Filtragem e Organização:

Filtragem de linhas inválidas, como registos sem ID, e filtragem específica para manter apenas filmes.

Reordenação dos dados com base no ID, mantendo uma sequência consistente para facilitar a visualização e interpretação.

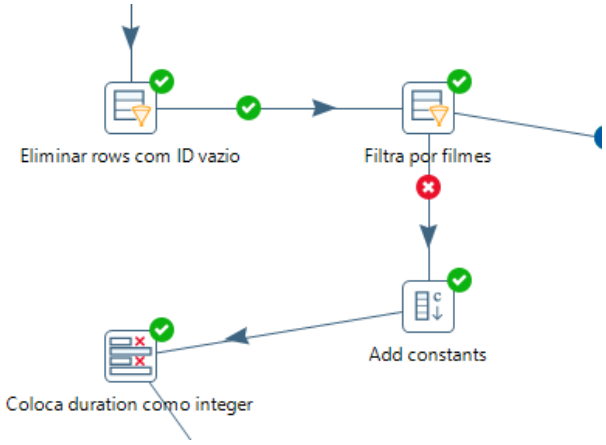

Figura 4 - Filtragem e Organização

Agregação e Enriquecimento:

União/cálculo dos campos de duração para obter uma visão geral do tempo total de cada registo.

Operações adicionais para agrupar as temporadas e episódios, para facilitar o entendimento do conteúdo no conjunto de dados.

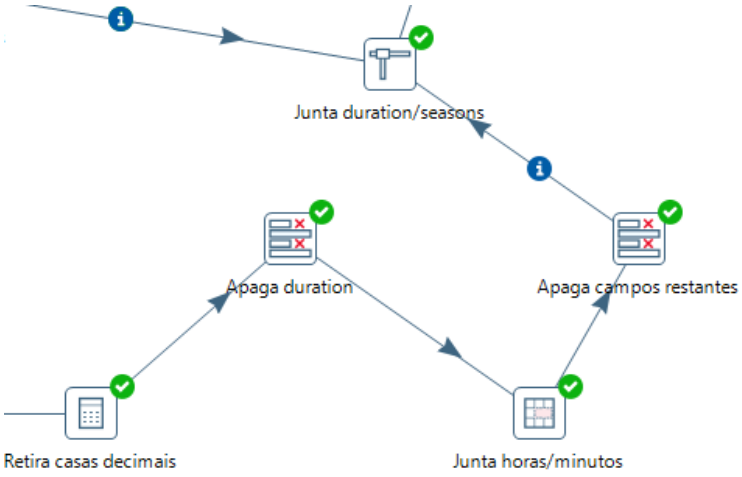

Figura 5 - Agregação e Enriquecimento

Import_Export

O fluxo de trabalho foi estruturado para integrar dados locais (de um arquivo CSV) com dados extraídos de uma API (através de JSON) e consolidar todas as informações de forma padronizada.

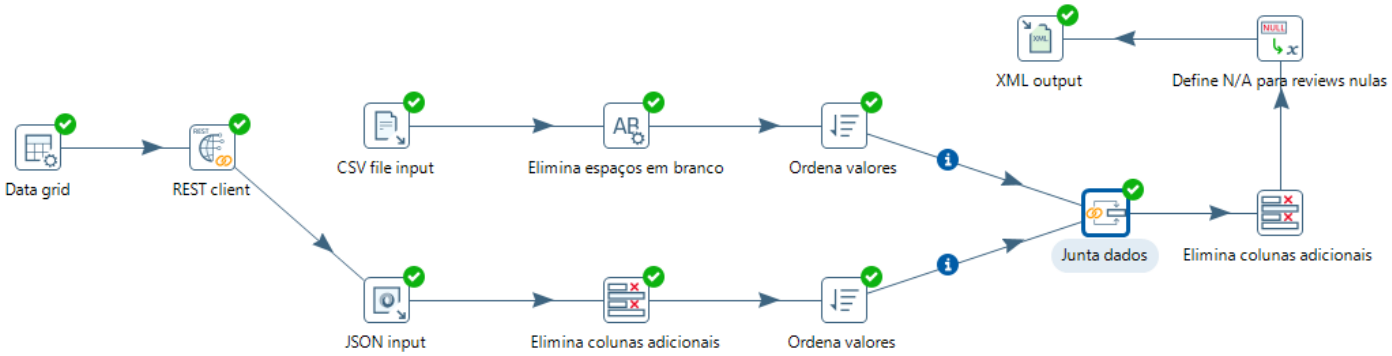

Figura 6 - Import_Export.ktr

Passos Principais da Transformação:

Extração de Dados:

REST Client: Conexão com a API para obter dados de filmes, os quais são armazenados em JSON.

Data Grid e CSV File Input: Carregamento dos dados originais da Netflix a partir de um arquivo CSV, permitindo uma comparação e enriquecimento posterior com os dados da API.

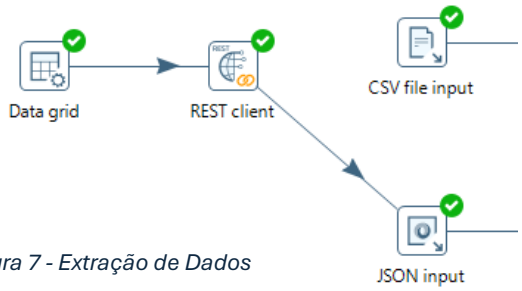

Figura 7 - Extração de Dados

Limpeza e Preparação de Dados:

Elimina Espaços em Branco: Remoção de espaços extras para garantir que os dados estejam uniformes e prontos para junção.

Elimina Colunas Adicionais: Exclusão das colunas desnecessárias nos dados da API, mantendo apenas os campos essenciais para o processo.

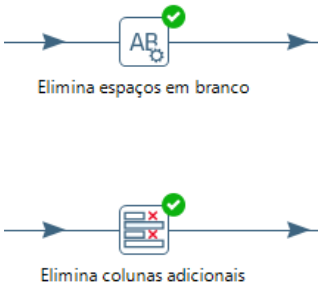

Figura 8 - Limpeza e Preparação

Ordenação e Padronização:

Ordena Valores: Alinhamento dos registos de ambos os conjuntos de dados (CSV e JSON) para facilitar o processo de junção.

Define N/A para Valores Nulos: Substituição de valores nulos por "N/A" para campos críticos como as avaliações dos filmes, garantindo consistência nas saídas.

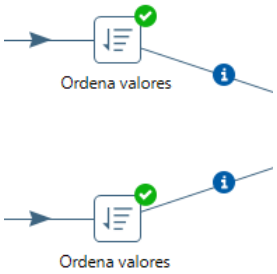

Figura 9 - Ordenação e Padronização

Integração e Consolidação:

Juntar Dados: Realização de um *Join* (junção) entre os dados do CSV e da API, combinando as informações por campos comuns (título). Esta junção permite enriquecer o conjunto de dados da Netflix com detalhes adicionais oriundos da API.

Elimina Colunas Adicionais: Removemos os campos que ficaram redundantes depois da junção, mantendo apenas as colunas finais e consistentes.

XML Output: Exportação dos dados consolidados em formato XML, permitindo que os dados enriquecidos estejam disponíveis para integração ou análise em sistemas externos.

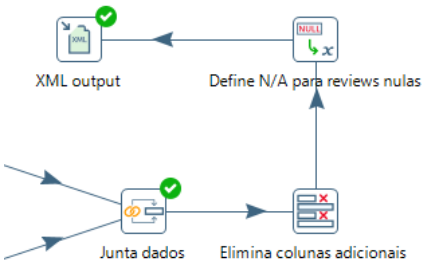

Figura 10 - Integração e Consolidação

HTMLOutput

Após a criação do ficheiro HTML, esta transformação é realizada para agregar as informações e configurar variáveis de ambiente com base nos dados de entrada.

Este processo complementa a etapa anterior, onde o arquivo HTML foi gerado com os dados já processados e formatados. A seguir, cada passo desta transformação é detalhado:

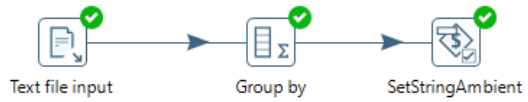

Figura 11 - HTMLOutput

Passos da Transformação

Leitura de Dados:

Text File Input: Este processo faz a leitura dos dados a partir de um ficheiro HTML previamente processado. Esse ficheiro contém os dados que foram utilizados e que agora serão agrupados para uma análise posterior ou para configurar variáveis que podem ser reutilizadas.

Agrupamento de Dados:

Com o step *Group by*, estamos a consolidar o conteúdo HTML num único campo, nomeado "html". O método utilizado é a concatenação de todas as linhas, permitindo que todo o conteúdo HTML seja agrupado sem separadores, resultando num único bloco de texto.

Configuração de Variáveis de Ambiente:

No passo *Set variables* (ou *SetStringAmbient*), o conteúdo da coluna html é definido como uma variável de ambiente com o nome *html*_body.

Esta variável está configurada para ter um alcance global, o que significa que pode ser utilizada noutras partes do processo, especialmente no envio de e-mails.

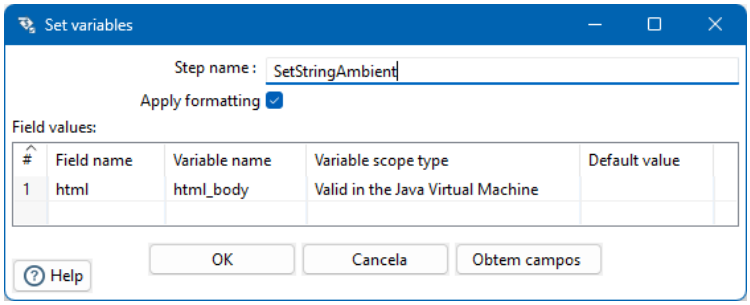

Figura 12 - SetStringAmbient

Jobs

ISI_Job

Este job no Pentaho Data Integration (*Spoon*) organiza um fluxo de operações ETL, onde os dados da API são processados, transformados em XML, convertidos para HTML, e enviados por e-mail. Aqui está uma explicação detalhada do processo:

Figura 13 - ISI_Job

Start: O ponto inicial do job. Ele inicia o processo completo e controla a sequência das tarefas.

Operações de dados e ER: Este passo realiza as operações de manipulação e limpeza dos dados recebidos, aplicando as expressões regulares (ER) para tratar e padronizar os dados. Esta etapa é importante para garantir que os dados estejam num formato consistente antes da exportação.

API e Export XML: Conecta-se a uma API para extrair dados em JSON. Após a extração, esses valores são agregados no *input* de dados original, e são exportados para um formato XML, que será utilizado nas etapas seguintes.

XML Existe: Este passo verifica se o ficheiro XML foi gerado com sucesso. Caso o ficheiro exista, o processo continua, caso contrário, é interrompido e gera uma mensagem de erro.

XSL Transformation: Converte o arquivo XML gerado para um formato de apresentação, como HTML, usando uma folha de estilo XSL (Extensible Stylesheet Language). Esta transformação é essencial para criar um HTML bem formatado para o envio.

Prepara Ficheiro HTML: Organiza e finaliza o arquivo HTML para garantir que o conteúdo esteja pronto para ser enviado. Aqui foram aplicados os ajustes finais na estrutura ou conteúdo do HTML.

Mail: Envia o arquivo HTML por e-mail. Nesta etapa, a variável html_body (criada anteriormente) é usada para inserir o conteúdo HTML como corpo do e-mail. Esse envio permite enviar os dados extraídos e transformados de forma organizada e visualmente apresentável.

Video Demonstração

Código QR - Vídeo

Conclusão

Em resumo, o trabalho alcançou com sucesso os objetivos definidos, mostrando a importância e o valor de processos de ETL bem estruturados e configurados para facilitar a integração e análise de dados em ambientes complexos.

O uso de expressões regulares, agregações e transformações com variáveis permitiu a criação de um fluxo de dados automatizado, que enriquece e organizou as informações de uma forma estruturada e padronizada.

Este processo também demonstrou a importância de uma abordagem planeada e com detalhe, considerando as particularidades de cada tipo de dados e as saídas necessárias para atender às exigências de sistemas de informação modernos.

O projeto demonstrou a eficácia das ferramentas como o Pentaho (*PDI*) no tratamento de dados, realizando várias operações que vão desde a normalização e limpeza de dados até à integração e consolidação em formatos variados.

Webgrafia

E-Learning (ISI) - https://elearning.ipca.pt/2425/course/view.php?id=34116