Instituto Politécnico do Cávado e do Ave

Escola Superior de Tecnologia

Licenciatura

em

Engenharia de Sistemas Informáticos

Trabalho Prático 1

Fábio Alexandre Gomes Fernandes – a22996

Barcelos, fevereiro de 2025

Instituto Politécnico do Cávado e do Ave

Escola Superior de Tecnologia

Licenciatura

em

Engenharia de Sistemas Informáticos

Trabalho Prático 1

Fábio Alexandre Gomes Fernandes – a22996

Unidade Curricular: Integração de Sistemas de Informação

Docente: Óscar Rafael da Silva Ferreira Ribeiro

Barcelos, maio de 2025

Ficha de Identificação

Elaborado por	Fábio Alexandre Gomes Fernandes
Número Mecanográfico	a22996

Unidade Curricular	Integração de Sistemas de Informação - ISI
Curso	Licenciatura em Engenharia de Sistemas Informáticos
Instituição	Escola Superior de Tecnologia do Instituto Politécnico do Cávado e do Ave
Professor Docente	Professor Óscar Rafael da Silva Ferreira Ribeiro
Contato	oribeiro@ipca.pt

Data Início	26 de setembro de 2025
Data de Conclusão	19 de outubro de 2025

Resumo

O presente trabalho foi desenvolvido no âmbito da unidade curricular de Integração de Sistemas de Informação (ISI), com o objetivo de aplicar de forma prática os conceitos de ETL (*Extract*, *Transform*, *Load*), recorrendo à ferramenta Pentaho Data Integration (PDI). O tema abordado, "Estatística de Streaming (Netflix)", permitiu explorar a integração e análise de dados relacionados com utilizadores e atividades de visualização de uma plataforma de *streaming*.

O projeto consistiu na criação de um *pipeline* ETL completo, englobando as fases de extração, transformação e carregamento de dados provenientes de dois conjuntos distintos: um relativo a utilizadores (com informações pessoais, tipo de subscrição e preferências) e outro referente à atividade de *streaming* (com dados de visualização, dispositivo e país).

Ao longo do desenvolvimento, procurou-se implementar todos os critérios de mais-valia estabelecidos no enunciado. Entre eles, destacam-se a utilização de expressões regulares (Regex) para normalização de campos, processos de limpeza e validação de dados, operações de junção e agregação, exportação de resultados em múltiplos formatos (CSV e JSON) e a criação de um *job* automatizado com *logs* e notificações por e-mail, assegurando o controlo e a rastreabilidade da execução.

O resultado evidencia a transformação de dados brutos e inconsistentes em informação estruturada e útil, demonstrando a capacidade de aplicar metodologias de integração de dados.

Índice

Índice de Figuras

1. Introdução

1.1 Contextualização

No âmbito da unidade curricular Integração de Sistemas de Informação (ISI), foi-nos proposto o desenvolvimento de um trabalho prático individual com o objetivo de aplicar os conceitos teóricos abordados na unidade, nomeadamente no que respeita aos processos de ETL (*Extract*, *Transform*, *Load*) e à integração de dados provenientes de diferentes fontes.

O trabalho é orientado pelo docente Dr. Óscar Rafael da Silva Ferreira Ribeiro e visa demonstrar a capacidade de planear, implementar e documentar um projeto completo de integração de dados, recorrendo a ferramentas adequadas ao contexto da unidade curricular.

1.2 Pretensões e Objetivos

O presente trabalho tem como principal pretensão demonstrar a aplicação prática de processos ETL (*Extract*, *Transform*, *Load*) através da análise e integração de dados relacionados com a plataforma de *streaming* Netflix.

Pretende-se extrair, limpar, transformar e combinar dados de diferentes origens, nomeadamente informação de utilizadores e de atividade de visualização. De modo a produzir um conjunto de dados unificado e fiável, capaz de suportar análises estatísticas e a geração de relatórios automatizados.

Os objetivos específicos incluem:

- Explorar e aplicar técnicas de limpeza, normalização e enriquecimento de dados;
- Implementar operações de junção (*stream lookups*), agrupamento e agregação para extração de métricas relevantes;
- Demonstrar o uso de expressões regulares, conversões de tipos e operações de validação;
- Realizar a exportação dos resultados em diferentes formatos (CSV e JSON);
- Implementar um *Job* de controlo e monitorização, incluindo *logs* de execução e notificações por e-mail;

1.3 Estrutura de Organização

Este relatório foi elaborado com o objetivo de apresentar, de forma clara e sequencial, o desenvolvimento do trabalho prático realizado no âmbito da unidade curricular Integração de Sistemas de Informação (ISI), cujo tema é "Estatística de Streaming (Netflix)". A estrutura do documento encontra-se organizada da seguinte forma:

- Capa e Ficha Técnica, contendo a identificação do aluno, unidade curricular, docente e instituição;
- Resumo, apresentando de forma sintética os objetivos, metodologia e resultados alcançados;
- Índices, com o Índice Geral, Índice de Figuras e Tabelas, e Lista de Siglas e Acrónimos;
- Introdução, onde é apresentado o enquadramento do trabalho, os objetivos propostos e a estrutura geral do relatório;
- Introdução ao Problema Abordado, com a descrição do tema "Estatística de Streaming", a sua relevância e o contexto em que o problema foi explorado;
- Tecnologias utilizadas, onde se descrevem as ferramentas de desenvolvimento aplicadas;
- Análise e Desenvolvimento, que detalha a implementação do processo ETL, incluindo a explicação das Transformações, dos Jobs de controlo e automação, e dos processos de exportação de dados;
- Demonstração, onde é apresentado o vídeo que ilustra o funcionamento completo do sistema e a execução dos processos desenvolvidos;
- Conclusão, que reflete sobre os resultados obtidos, as competências adquiridas e as possíveis melhorias futuras;
- Referências Bibliográficas, reunindo as fontes e recursos consultados durante o desenvolvimento do projeto.

2. Análise do Problema

2.1 Descrição do Problema

O presente trabalho tem como principal foco o desenvolvimento de um processo de integração, transformação e análise de dados (ETL) aplicado a um contexto de plataformas de *streaming* digital, tomando como caso de estudo o serviço Netflix.

O problema identificado centra-se na necessidade de recolher, limpar, normalizar e relacionar grandes volumes de dados provenientes de diferentes fontes, de forma a permitir a obtenção de informação estruturada, coerente e analiticamente útil.

Na realidade das plataformas de *streaming*, os dados recolhidos diariamente, como o tempo de visualização, o tipo de dispositivo utilizado, o país de origem do utilizador, o género de conteúdos preferido ou o tipo de subscrição, são fundamentais para a tomada de decisões estratégicas. Contudo, esses dados encontram-se frequentemente dispersos, redundantes ou em formatos heterogéneos, dificultando a análise direta.

Neste contexto, o trabalho propõe-se a simular e resolver esse problema através da criação de um *pipeline* ETL completo, que integra dois conjuntos de dados:

- Um *dataset* de utilizadores da Netflix, com informações demográficas, tipo de subscrição e preferências de visualização;
- Um *dataset* de atividades de *streaming*, contendo o histórico de visualizações, dispositivos utilizados e localização geográfica.

O desafio consiste, portanto, em construir um processo automatizado que leia ambos os *datasets*, realize operações de limpeza, transformação e junção, e produza um conjunto final de dados enriquecidos e prontos para análise.

2.2 Fontes de Dados (*Datasets*)

O processo de ETL desenvolvido opera sobre dois *datasets* de origem distintos, que simulam fontes de dados heterogéneas com diferentes desafios de qualidade e estrutura.

- **Dataset 1:** "*netflix_activity.csv*"
  - **Descrição:** Este ficheiro contém os registos detalhados da atividade de visualização na plataforma. Cada linha representa um evento de visualização único, contendo informação sobre o que foi visto, quando e como.
  - **Campos:** *Profile Name*, *Start Time*, *Duration*, *Attributes*, *Title*, *Supplemental Video Type*, *Device Type*, *Bookmark*, *Latest Bookmark*, *Country*.
  - **Desafios:** Os principais desafios deste *dataset* são de natureza estrutural e de formato. Os dados, embora consistentes, não estão prontos para análise. Os problemas incluem nomes de colunas com espaços, dados de tempo em formato de texto (HH:MM:SS) e informação de país agregada (PT (Portugal)) que necessita de ser extraída e separada.

- **Dataset 2:** "*netflix_users.csv*"
  - **Descrição:** Este ficheiro contém os dados demográficos e de subscrição dos utilizadores da plataforma. Para o propósito deste trabalho, foi concebido com múltiplos erros para simular um cenário real de "dados sujos" que necessitam de um processo de limpeza, validação e correção. Serviu também de base para a criação de uma fonte de dados em formato JSON, para demonstrar a importação de múltiplos formatos.
  - **Campos:** *User_ID*, *Profile_Name*, *Real_Name*, *Age*, *Country*, *Subscription_Type*, *Watch_Time_Hours*, *Favorite_Genre*, *Last_Login*.

- **Desafios:** Ao contrário do primeiro *dataset*, os desafios aqui são de qualidade e integridade dos dados, simulando erros de inserção manual. Os problemas incluem valores inválidos, erros de escrita em campos categóricos (ex: *Prremium*, *Satndar*, *sabic*), dados em falta e formato inconsistente.

2.3 Tecnologias utilizadas

Para o desenvolvimento deste projeto, foram selecionadas as seguintes tecnologias e ferramentas, com o objetivo de construir uma solução de ETL robusta, visual e alinhada com as práticas do mercado de integração de dados.

- ***Pentaho Data Integration* (PDI) 10.2:** Foi a ferramenta principal de ETL (*Extract*, *Transform*, *Load*) utilizada para desenhar, executar e orquestrar todo o processo. A sua natureza visual, baseada em fluxos de dados, permitiu a rápida prototipagem, depuração e implementação de regras complexas de limpeza, transformação e validação de dados. A distinção entre Transformações (para a manipulação de dados) e *Jobs* (para a orquestração e controlo de fluxo) foi fundamental para a arquitetura da solução.

- ***Git*:** O sistema de controlo de versões distribuído foi utilizado para gerir o histórico de todo o material produzido, incluindo os ficheiros de projeto do Pentaho (.ktr, .kjb), os *datasets* e a própria documentação. A sua utilização garantiu a integridade e o rastreamento de todas as alterações efetuadas ao longo do desenvolvimento. O repositório completo pode ser acedido em: https://github.com/fabiofernandes6/TP01-de-ISI.git

- ***GitHub*:** A plataforma de alojamento baseada na web foi usada para hospedar o repositório Git do projeto. Serviu como um ponto central para a partilha, documentação e versionamento do código-fonte da solução, cumprindo os requisitos de entrega e boas práticas de desenvolvimento de *software*.

- **Formatos de Dados (CSV e JSON):** Parte integral da estratégia tecnológica foi a capacidade de lidar com múltiplos formatos de representação de dados. O processo demonstrou a importação de dados de ficheiros CSV, e a exportação dos resultados para CSV e JSON, evidenciando a flexibilidade da solução para se integrar com diferentes sistemas.

3. Análise e Desenvolvimento de Software

A resolução do problema descrito no capítulo anterior foi implementada através de um processo de ETL, utilizando a ferramenta *Pentaho Data Integration*, também conhecida como *Spoon*. A estratégia de desenvolvimento seguiu uma abordagem modular, separando as responsabilidades em duas componentes distintas e complementares:

- Uma **Transformação** (.ktr), que constitui o núcleo do processo, onde toda a lógica de extração, limpeza, validação, enriquecimento e carregamento dos dados é executada.

- Um **Job** (.kjb), que atua como o orquestrador do processo, responsável por gerir o fluxo de execução, tratar erros, garantir pré-condições e controlar as notificações.

As secções seguintes detalham a implementação técnica de cada uma destas componentes, desde o fluxo de dados na transformação até à lógica de controlo no Job.

3.1 Transformação (.ktr)

A transformação "***streaming_netflix.ktr***" é o coração do processo de ETL. Foi desenhada para extrair os dados das fontes heterogéneas, aplicar um *pipeline* de limpeza e validação robusto, enriquecer os dados através da sua fusão, e finalmente, carregar os resultados em múltiplos formatos. A sua implementação é detalhada nas secções seguintes.

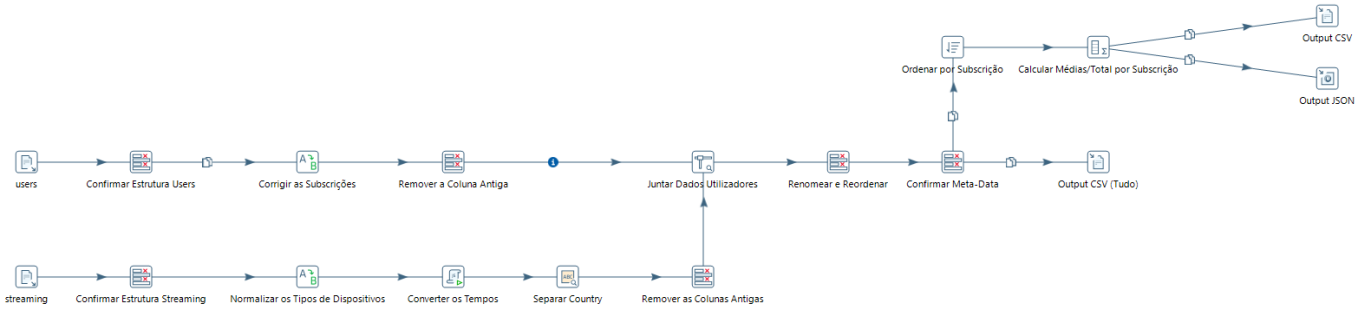

Figura 1: Transformação "streaming_netflix.ktr"

3.1.1 Extração das Fontes de Dados

O processo inicia-se com a extração de dados de duas fontes distintas, que correm em fluxos paralelos:

- **Fluxo de Atividades:** Utiliza o *step* "*CSV file input*" para ler os dados de atividade do ficheiro "*netflix_activity.csv*".
- **Fluxo de Utilizadores:** Utiliza o *step* "*CSV file input*" para ler os dados demográficos e de subscrição do ficheiro "*netflix_users.csv*", que contém múltiplos erros de qualidade de dados.

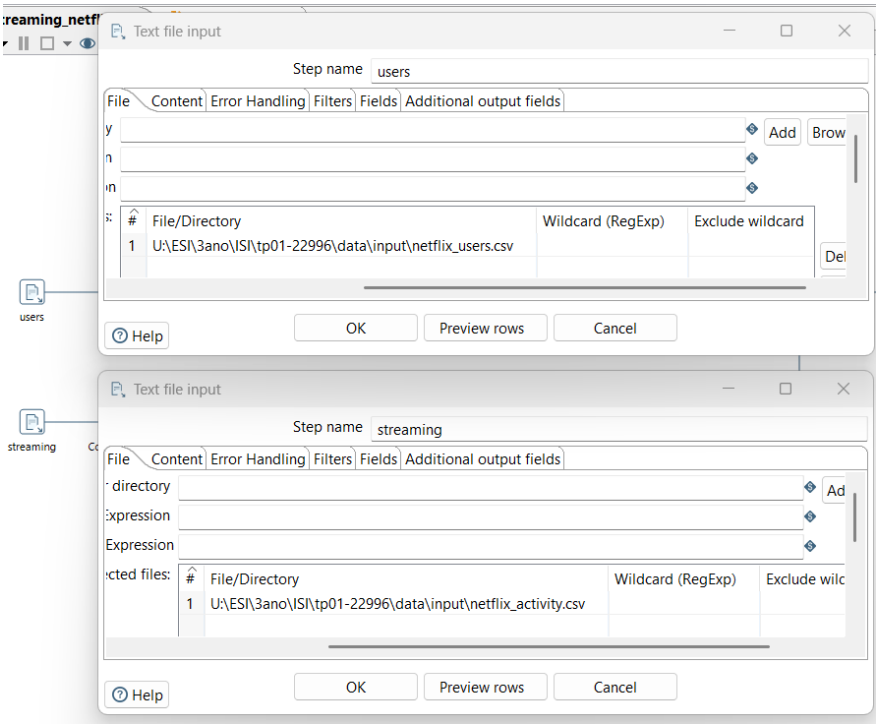

Figura 2: Extração das fontes de dados CSV

3.1.2 Fluxo de Utilizadores

Este fluxo é focado na confirmação, limpeza e correção intensiva dos dados de utilizadores, aplicando uma sequência de passos para garantir a máxima qualidade dos dados antes do *join*. O processamento contou com os seguintes passos:

1. **Confirmar Estrutura *Users*:** Utiliza um *step* "*Select Values*". Este passo serve para confirmar os campos do ficheiro original que contém os dados dos utilizadores;
2. **Corrigir as Subscrições:** Utiliza um *step* "*Value Mapper*". Este passo serve para corrigir os erros de escrita específicos no campo "*Subscription_Type*" (ex: "*Prremium*" = "*Premium*");
3. **Remover a Coluna Antiga:** Utiliza um *step* "*Select Values*". Este passo serve para remover o antigo campo "*Subscription_Type*" pelo novo "*Subscription_Type_Corrected*" onde contém a correção dos erros de escrita.

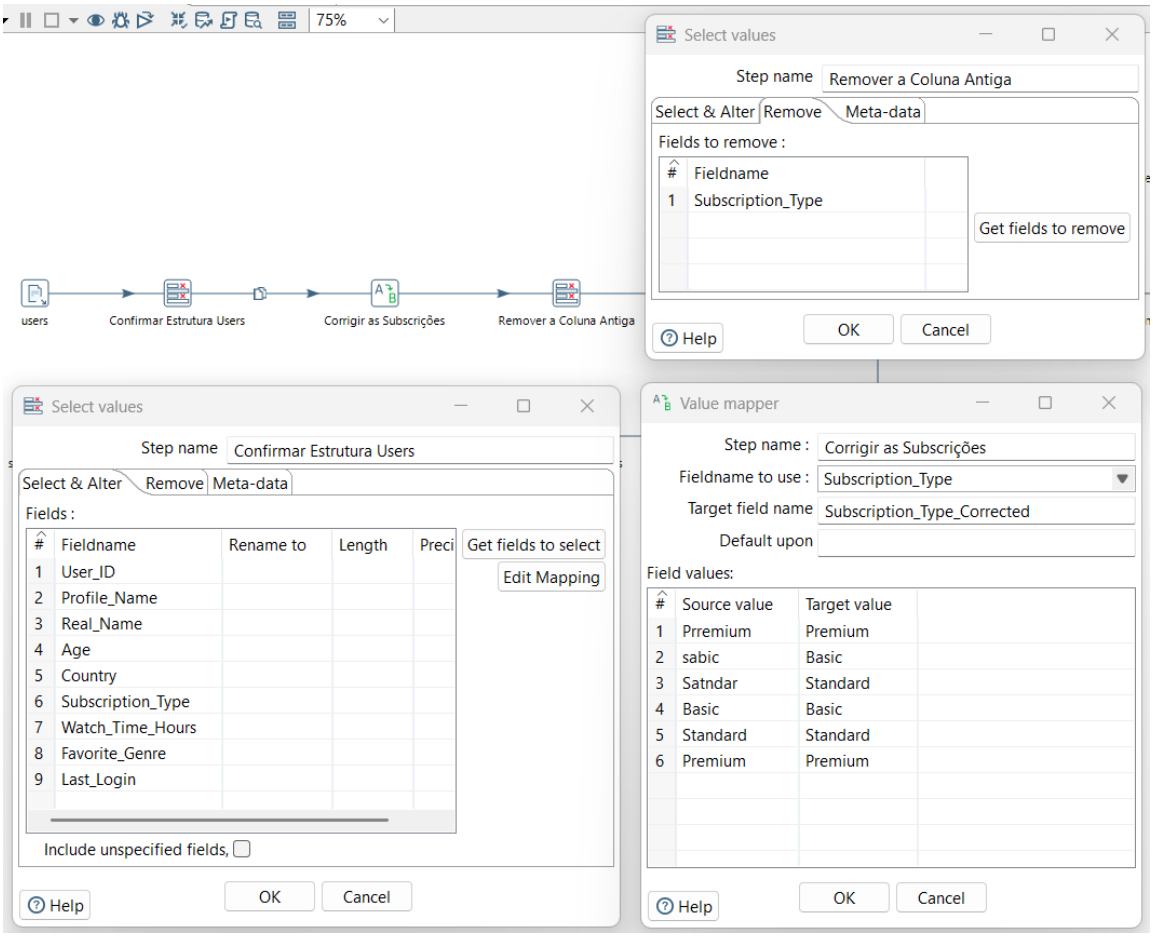

Figura 3: Fluxo de Utilizadores

3.1.3 Fluxo de Atividades

Este fluxo é responsável pela limpeza, normalização e preparação dos dados de atividades de visualização, garantindo que todos os campos estejam prontos para integração com o fluxo de utilizadores. O processamento foi realizado através dos seguintes passos:

1. **Confirmar Estrutura *Streaming*:** Utiliza um *step* "*Select Values*". Este passo serve para confirmar a estrutura e os campos do ficheiro original que contém os dados de atividades de visualização;
2. **Normalizar os Tipos de Dispositivos:** Utiliza um *step* "*Value Mapper*". Serve para normalizar os valores do campo "*Device_Type*", convertendo identificadores numéricos em categorias textuais mais legíveis (ex.: "*Device Type 1*" = "*Computer*");
3. **Converter os Tempos:** Utiliza um *step* "*Modified JavaScript Value*". Este passo converte o tempo da duração da atividade (campo "*Duration*") do formato "hh:mm:ss" para o total de horas decimais, criando um novo campo convertido.

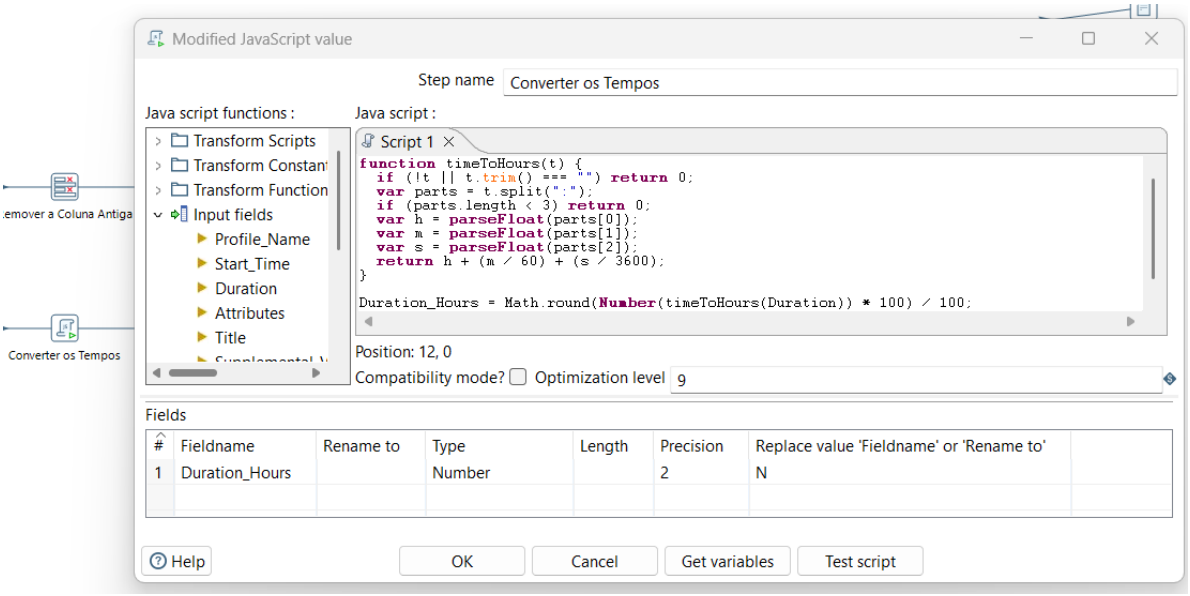

Figura 4: Fluxo de Atividades - Converter os Tempos

4. **Separar Country:** Utiliza um *step* "*Regex Evaluation*". Este passo aplica uma expressão regular ao campo "*Country*" para separar o código do país e o nome do país em dois novos campos distintos ("*Country_Code*" e "*Country_Name*").

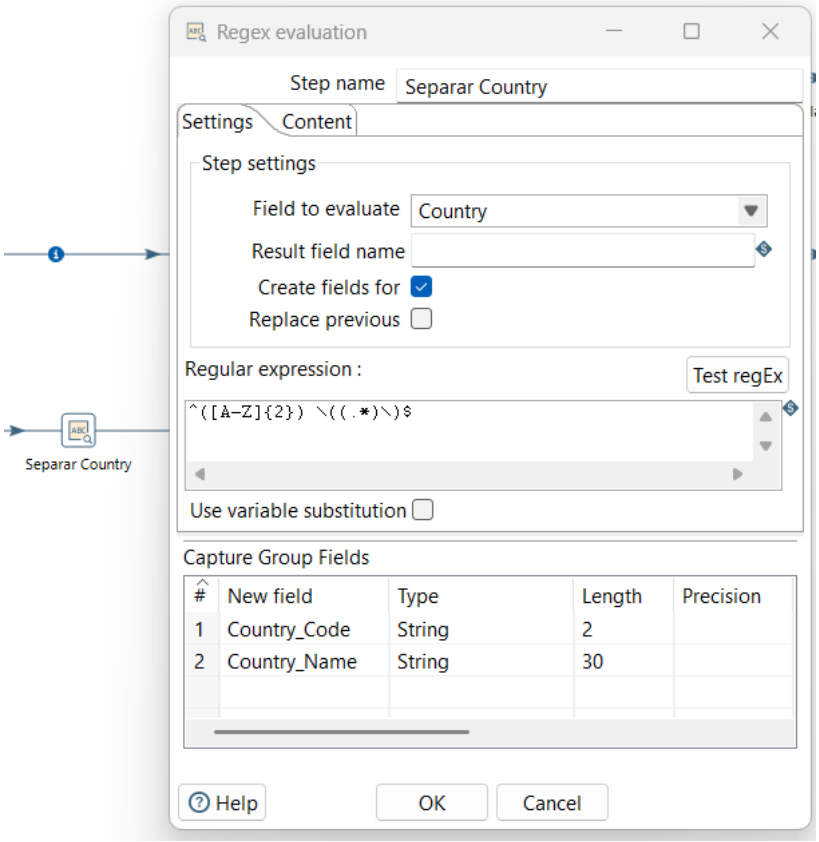

Figura 5: Fluxo de Atividades - Separar *Country*

5. **Remover as Colunas Antigas:** Utiliza um *step* "*Select Values*". Serve para eliminar os campos desnecessários antigos e temporários (como "*Device_Type*" original ou tempos não convertidos), mantendo apenas os campos normalizados e preparados para a integração com os dados de utilizadores.

3.1.4 Junção e Enriquecimento (Join)

Esta fase representa o ponto de convergência entre os fluxos de Atividades e Utilizadores, onde os dados previamente tratados são integrados num único dataset consolidado.

1. **Juntar Dados Utilizadores:** Utiliza um *step* "*Stream Lookup*".O fluxo principal, proveniente das atividades, é enriquecido com as informações já limpas e normalizadas do fluxo de utilizadores. O *step* "*Stream Lookup*" utiliza o campo "*Profile_Name*" como chave de junção, garantindo a correspondência entre o perfil de cada atividade e os respetivos dados do utilizador (como nome real, idade, país, tipo de subscrição, género favorito, horas já assistidas e o último *login*).

O resultado é um *dataset* completo e coeso, combinando dados comportamentais (atividades) e dados demográficos (utilizadores), pronto para análise e exportação.

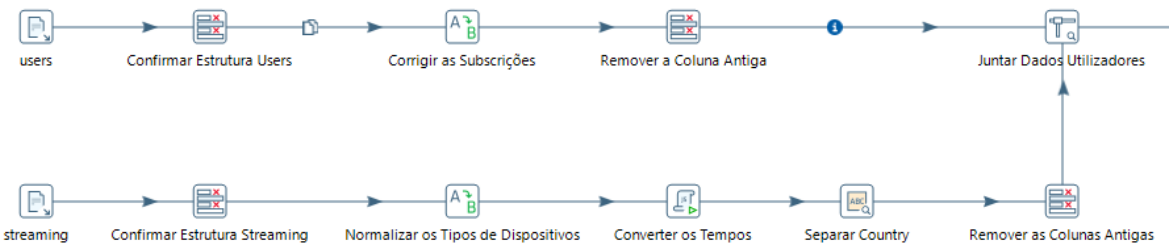

Figura 6: Junção e Enriquecimento (*Join*)

3.1.5 Agregação e Saídas (Load)

Após a junção dos dados dos utilizadores com as atividades, o fluxo é dividido em dois ramos distintos para gerar tanto os dados detalhados como os relatórios agregados.

- **Ramo Principal - Saída Detalhada:**
  1. O fluxo proveniente do "*Stream Lookup*" é encaminhado para o *step* "Renomear e Reordenar", que organiza e uniformiza os nomes das colunas resultantes do processo de junção.

2. De seguida, o *step* "Confirmar Meta-Data" garante a consistência e o tipo de dados de cada campo antes da exportação.
3. Por fim, o "*Output* CSV (Tudo)" grava o *dataset* completo e enriquecido num ficheiro CSV, contendo todas as atividades já associadas às informações dos respetivos utilizadores.

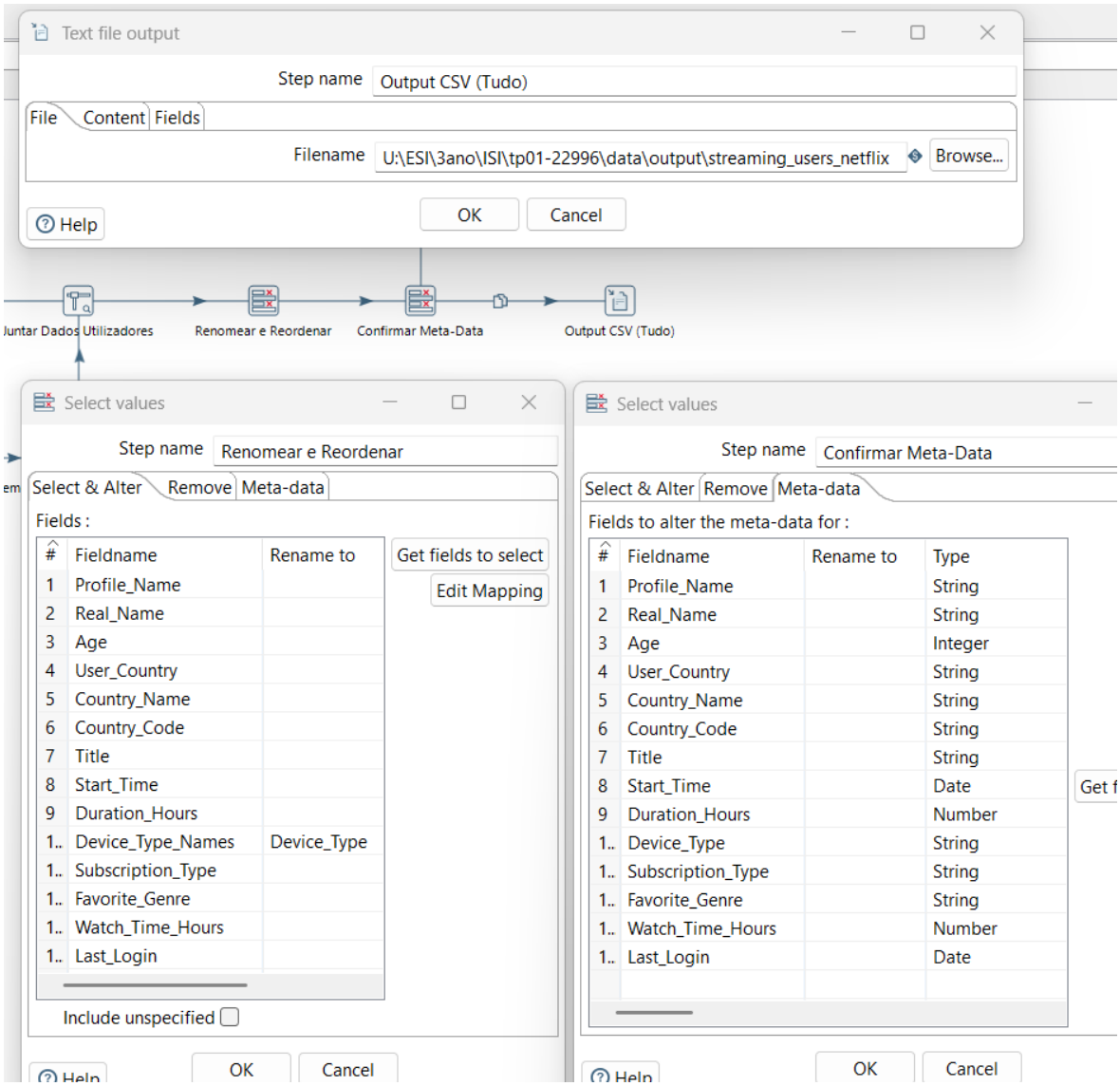

Figura 7: Agregação e Saídas (*Load*) - Saída detalhada

- **Ramo Secundário – Saída Agregada:**
  1. A partir da estrutura consolidada, o fluxo é desviado para um *step* "Ordenar por Subscrição", "*Sort rows*", que organiza os registos por tipo de plano.
  2. Em seguida, o *step* "Calcular Médias/Total por Subscrição", "*Group by*", aplica operações de agregação, calculando o total de horas, a média de horas visualizadas e o número de utilizadores por tipo de subscrição.
  3. Os resultados deste agrupamento são depois exportados em diferentes formatos, CSV e JSON, para facilitar a integração com outros sistemas e demonstrar a versatilidade da solução implementada.

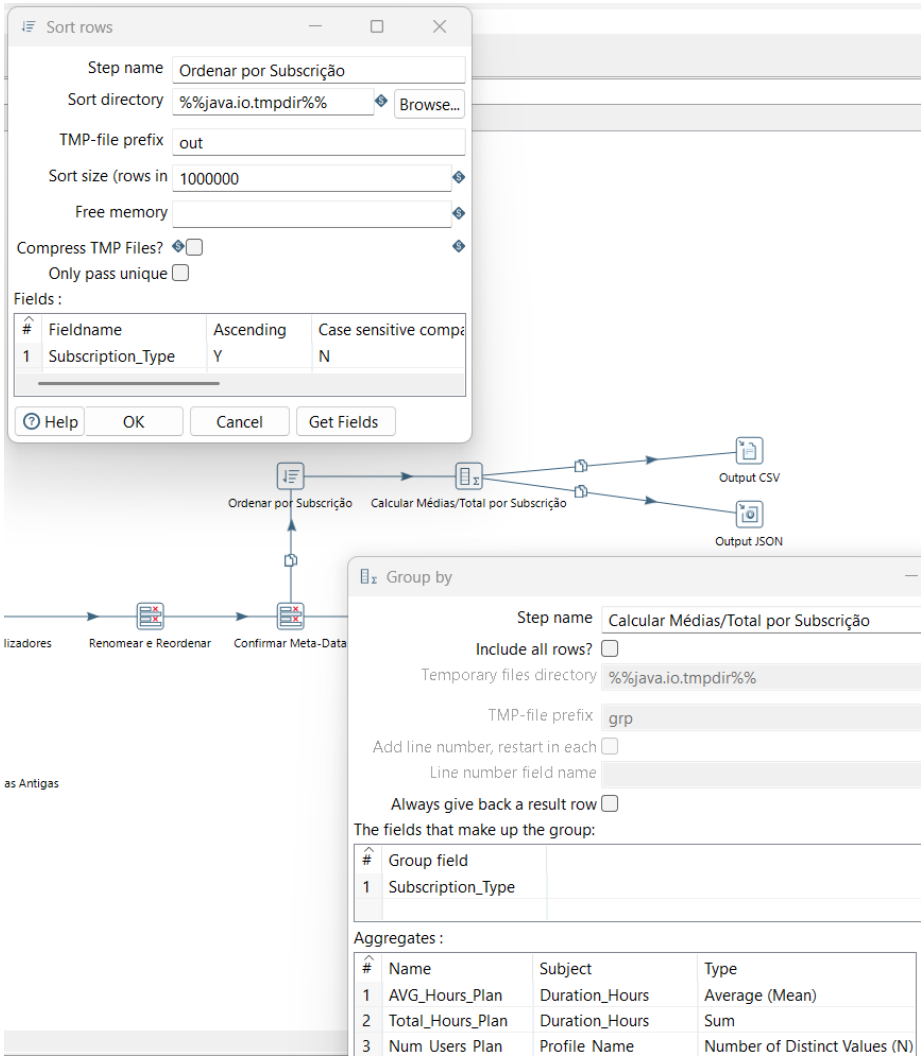

Figura 8: Agregação e Saídas (*Load*) - Saída agregada

3.2 *Job* (.kjb)

O *job* "*project_netflix*.kjb", ilustrado na figura 9, foi desenvolvido para automatizar e monitorizar todo o processo ETL, garantindo o controlo total sobre a execução da transformação principal (*streaming_netflix*.ktr), a verificação prévia dos ficheiros de entrada e a notificação automática por email consoante o resultado.

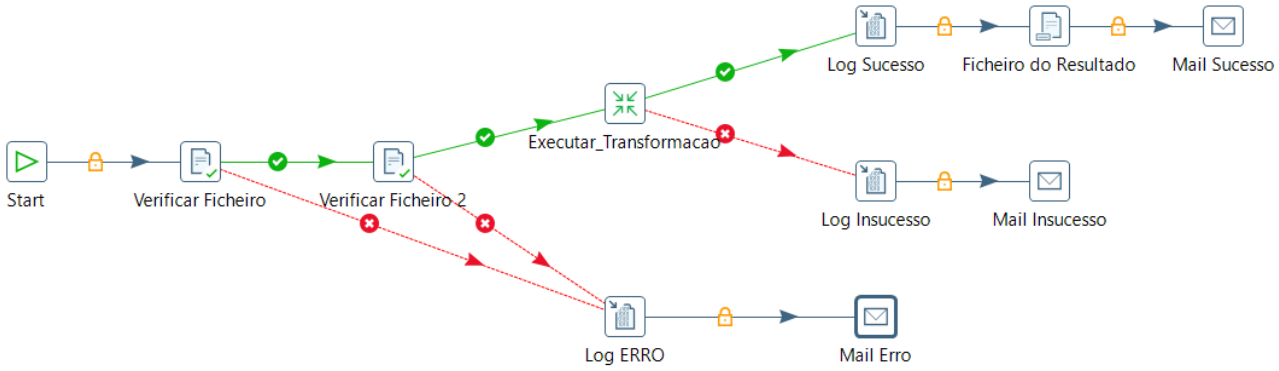

Figura 9: Job (.kjb)

O fluxo do job segue a seguinte lógica:

- ***Start*:** Inicia o processo e direciona a execução para os passos de verificação.

- **Verificar Ficheiros:** São utilizados dois *steps* "*File Exists*", um para confirmar a presença do ficheiro de atividades (*netflix_activity*.csv) e outro para o de utilizadores (*netflix_users*.csv), como é reparado na figura 10.
  - Se ambos os ficheiros existirem, o fluxo segue para a transformação principal.
  - Caso algum ficheiro esteja em falta, é ativado o ramo de erro, que gera um log (*step* "*Write to log*", "*Log* ERRO") específico e envia uma notificação por email (*step* "*Mail*", "*Mail* Erro").

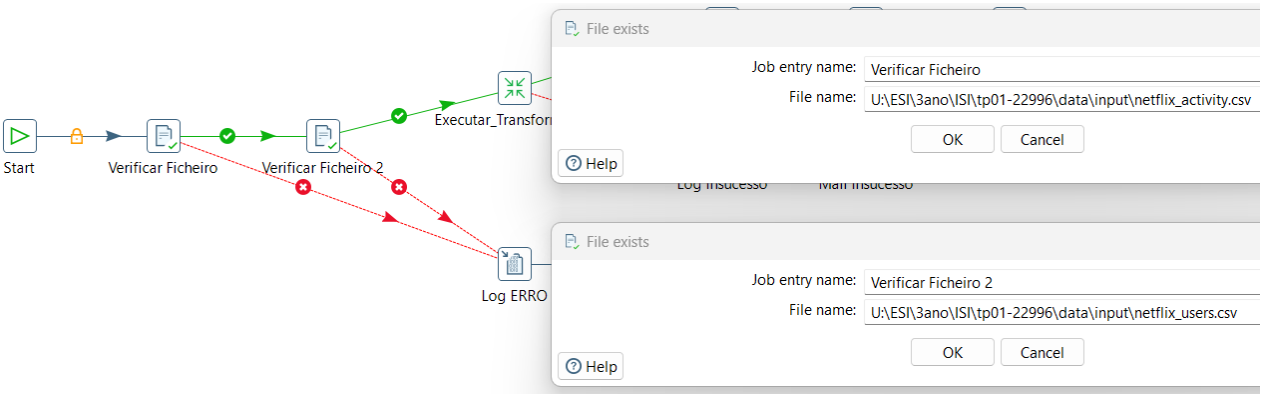

Figura 10: Verificação de ficheiros de entrada

- **Executar Transformação:** Este passo executa a transformação principal (*streaming_netflix*.ktr), responsável pelo tratamento, normalização, junção e exportação dos dados.

- **Gestão de Sucesso e Insucesso:**
  - "*Log* Sucesso" – "Ficheiro do Resultado" – "*Mail* Sucesso": Caso a execução da transformação decorra sem erros, é criado um *log* de sucesso ("*Write to log*")e enviados, por email ("Mail"), os ficheiros de saída (*dataset* enriquecido e relatórios de subscrição).
  - "Log Insucesso" – "Mail Insucesso": Se a transformação falhar, é criado um log de insucesso ("*Write to log*") e enviado um por *email* ("*Mail*") com os detalhes do erro.

Este job assegura:

- Validação prévia dos ficheiros antes da execução da transformação;
- Monitorização e registo detalhado de *logs*;
- Notificações automáticas via *email*, diferenciando sucesso, insucesso e erro de ficheiros;
- Envio automático dos relatórios e *datasets* gerados.

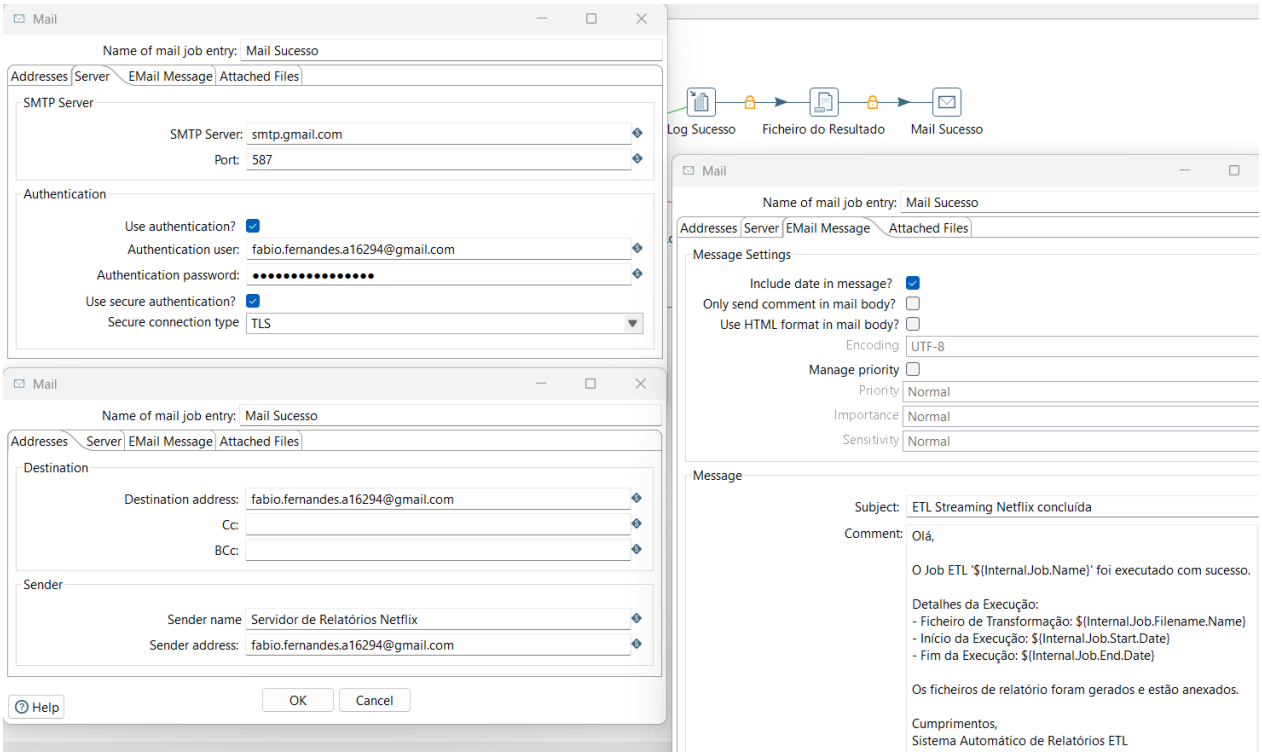

Figura 11: Configuração do *step* "*Mail* Sucesso"

3.3 Resultados

A execução completa da transformação e do job resultou na geração de dois tipos principais de outputs:

- ***Dataset* detalhado: "***streaming_users_netflix*.csv", que reúne e enriquece as atividades de visualização com as informações dos utilizadores;

A	B	C	D	E	F	G	H	I	J	K
Profile_Na	Real_Name	Start_Time	Duration_Hours	Title	Device_Typ	Country_C	Country_N	Subscriptic	Favorite_Genre	
User 1	Alex Johnson	28/04/2022 16:08	0	Chernobyl 1986	Computer	IN	India	Basic	Comedy	
User 1	Alex Johnson	28/04/2022 15:54	0,16	13 Hours: The Secret Soldiers of Benghazi	Computer	IN	India	Basic	Comedy	
User 1	Alex Johnson	28/04/2022 15:53	0	Saving Private Ryan	Computer	IN	India	Basic	Comedy	
User 1	Alex Johnson	28/04/2022 15:52	0	Mosul	Computer	IN	India	Basic	Comedy	

Figura 12: Dataset detalhado - "streaming_users_netflix.csv"

- **Relatório agregado:** "relatório_por_subscricao" (exportado em CSV e JSON), que apresenta a média de horas vistas, o total de horas e o número de utilizadores por tipo de subscrição.

	A	B	C	D
1	Subscription_Type	AVG_Hours_Plan	Total_Hours_Plan	Num_Users_Plan
2	Basic	0.26	1946.73	3
3	Premium	0.22	335.76	4
4	Standard	0.18	197.42	2

Figura 13: Relatório agregado: "relatório_por_subscricao.csv"

Ambos os ficheiros foram validados com sucesso e demonstram a correta execução do processo ETL, desde a limpeza e junção dos dados até à exportação final.

4. Vídeo de demonstração

A demonstração completa do projeto, desde a execução do *Job* no Pentaho até à visualização dos ficheiros de saída e das notificações por *email*, pode ser acedida através do seguinte QR Code.

Figura 14: QR Code do vídeo de demonstração

5. Conclusão

O desenvolvimento deste trabalho permitiu aplicar, de forma prática, os princípios fundamentais da integração de sistemas de informação, consolidando os conhecimentos adquiridos sobre processos ETL, manipulação de dados e orquestração de fluxos automatizados.

Através da ferramenta Pentaho Data Integration (PDI), foi possível construir uma solução completa, composta por transformações e um job de controlo, que demonstra todo o ciclo de integração, desde a leitura dos dados brutos até à geração e exportação de resultados consolidados.

Ao longo do projeto, foi dada especial atenção à implementação dos critérios de mais-valia definidos no enunciado. Foram exploradas expressões regulares (Regex) para normalização de campos, operações de limpeza e correção de dados, *joins* e agrupamentos para consolidação da informação, e a serialização em múltiplos formatos (CSV e JSON). Além disso, foi implementado um sistema de monitorização e controlo, com geração automática de *logs* e envio de notificações por *e-mail*, garantindo rastreabilidade e automatização do processo.

Estes aspetos permitiram demonstrar não apenas a capacidade de desenvolver um processo ETL funcional e eficiente, mas também de integrar diversos mecanismos complementares que aumentam a qualidade, robustez e valor analítico da solução.

Como trabalhos futuros, destaca-se a possibilidade de armazenar os dados tratados numa base de dados relacional, desenvolver *dashboards* interativos (em Power BI ou Apache Superset) e integrar dados provenientes de APIs externas, de forma a enriquecer a análise e expandir a utilidade do sistema.

Em suma, o projeto atingiu plenamente os objetivos propostos, evidenciando a aplicação prática dos conceitos de integração de sistemas e a tentativa de implementação de todos os critérios de mais-valia sugeridos, culminando num processo ETL completo, automatizado e alinhado com as boas práticas da área

6. Referências

GitHub Docs. (n.d.). Retrieved October 19, 2025, from https://docs.github.com/pt

JavaScript Guide - JavaScript | MDN. (n.d.). Retrieved October 19, 2025, from https://developer.mozilla.org/en-US/docs/Web/JavaScript/Guide

JSON. (n.d.). Retrieved October 19, 2025, from https://www.json.org/json-en.html

.NET Regular Expressions - .NET | Microsoft Learn. (n.d.). Retrieved October 19, 2025, from https://learn.microsoft.com/en-us/dotnet/standard/base-types/regular-expressions

Netflix Users Database. (n.d.). Retrieved October 19, 2025, from https://www.kaggle.com/datasets/smayanj/netflix-users-database?utm_source=chatgpt.com

Netflix Watch Log. (n.d.). Retrieved October 19, 2025, from https://www.kaggle.com/datasets/arjunajn/netflix-watch-log?utm_source=chatgpt.com