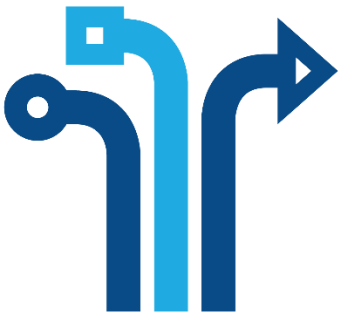

ESCOLA
SUPERIOR
DE TECNOLOGIA
IPCA

Tratamento de Dados com Recurso a Ferramenta de ETL

UTILIZAÇAO DO PENTAHO KETTLE PARA TRATAMENTO DE DADOS DE ARQUIVOS CSV

José Luís Rodrigues|17616 | ISI

Introdução

ÂMBITO

No âmbito da disciplina de Integração de Sistemas Informáticos pretende-se consolidar conceitos apresentados nas aulas ao utilizar as funcionalidades de aplicações de ETL sendo que a aplicação utilizada neste trabalho foi o Kettle.

PROBLEMA EM ESTUDO

Neste projeto os dados em estudo são filmes, tendo esta lista vindo das bases de dados do *IMBD,* e uma lista de series e filmes disponíveis na *Netflix*.

Arquitetura da solução

ESTRUTURA DA SOLUÇÃO

Inicialmente a solução iria ser desenvolvida em múltiplas fases de transformação mas devido a um problema de memoria que gerava o erro "java.lang.OutOfMemoryError: Java heap space" esta teve que ser desenvolvida numa só transformação.

DESCRIÇÃO DA SOLUÇÃO

Inputs

- Arquivo com os dados dos filmes em formato csv;
- Arquivo com os dados dos atores em formato csv;
- Arquivo com os filmes e series disponíveis na *Netflix* em formato csv;

Estes arquivos encontram-se na pasta "inputs".

Links para download dos arquivos:

- https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+title_principals.csv
- https://www.kaggle.com/shivamb/netflix-shows

Outputs

- Arquivo com os filmes de longas metragens (FeaturaFilms.json)
- Arquivo com os atores e os filmes em que participaram (Actors&Films.json)
- Arquivo com os atores que já morreram (DeathActors.json)
- Arquivo com os filmes comuns da Netflix e IMDB (Netflix&IMDB.json)

Estes arquivos encontram-se na pasta outputs.

Para o processo de listagem dos filmes em que cada ator participou, atores falecidos e longas metragens os ficheiros são carregados para o programa e sofrem as seguintes alterações:

- Atores
  - São removidas colunas com informação que não serão utilizadas;
  - Os dados são ordenados alfabeticamente pelo nome do ator

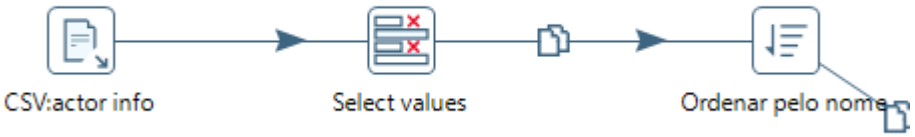

Figura 1 Modificações iniciais no arquivo dos atores

- Filmes
  - Os vídeos são separados por cada ator que participa neles, isto é, se num filme participarem 3 atores o resultado será uma lista com 3 linhas em que cada uma contem o nome do ator, nome e restante informação do filme;
  - São removidas colunas com informação que não serão utilizadas;
  - Os dados são ordenados pelo nome do ator;

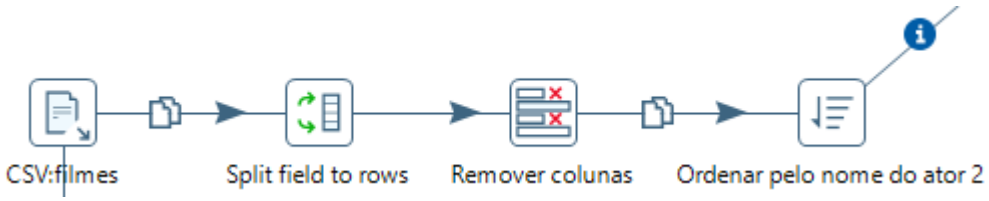

Figura 2 Modificações iniciais no arquivo dos filmes

Após estas primeiras alterações os dados são unidos utilizando o operador "*Merge join*" e devido a esta união existem dados repetidos que foram eliminados.

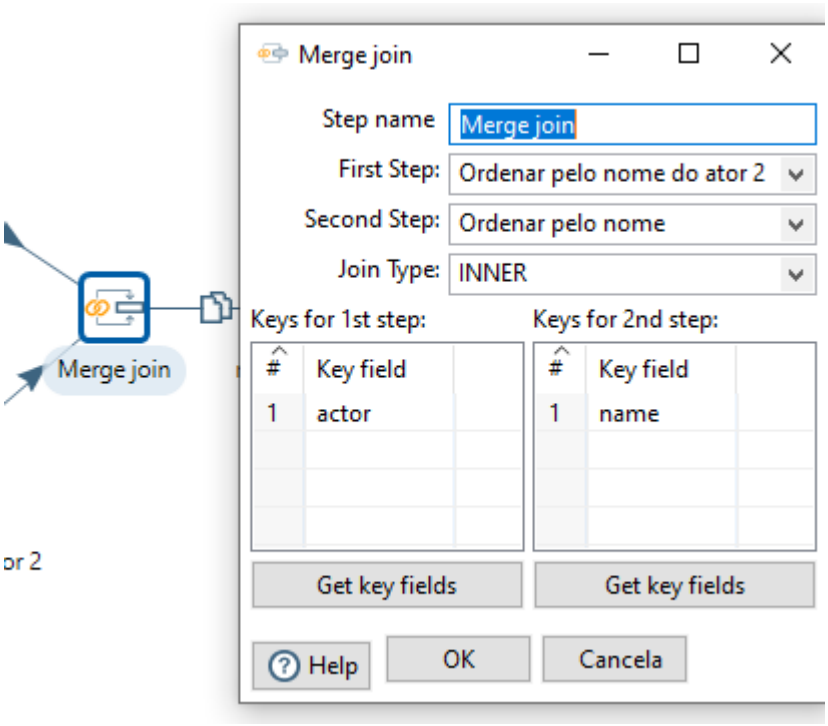

Figura 4 Configuração do Merge join

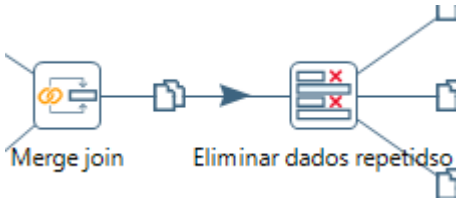

Figura 3 Merge join e remoção de dados repetidos

Após a eliminação dos dados repetidos são efetuadas as operações necessárias para que seja possível chegar a alguns dos objetivos finais.

- Listar os filmes de cada ator ordenados alfabeticamente.
  - São removidos dados não necessários, estes dados são ordenados pelo ator e é utilizado o operador "*Group by*" para agrupar os dados de cada ator corretamente.

Figura 5 Processo completo para a listagem dos filmes

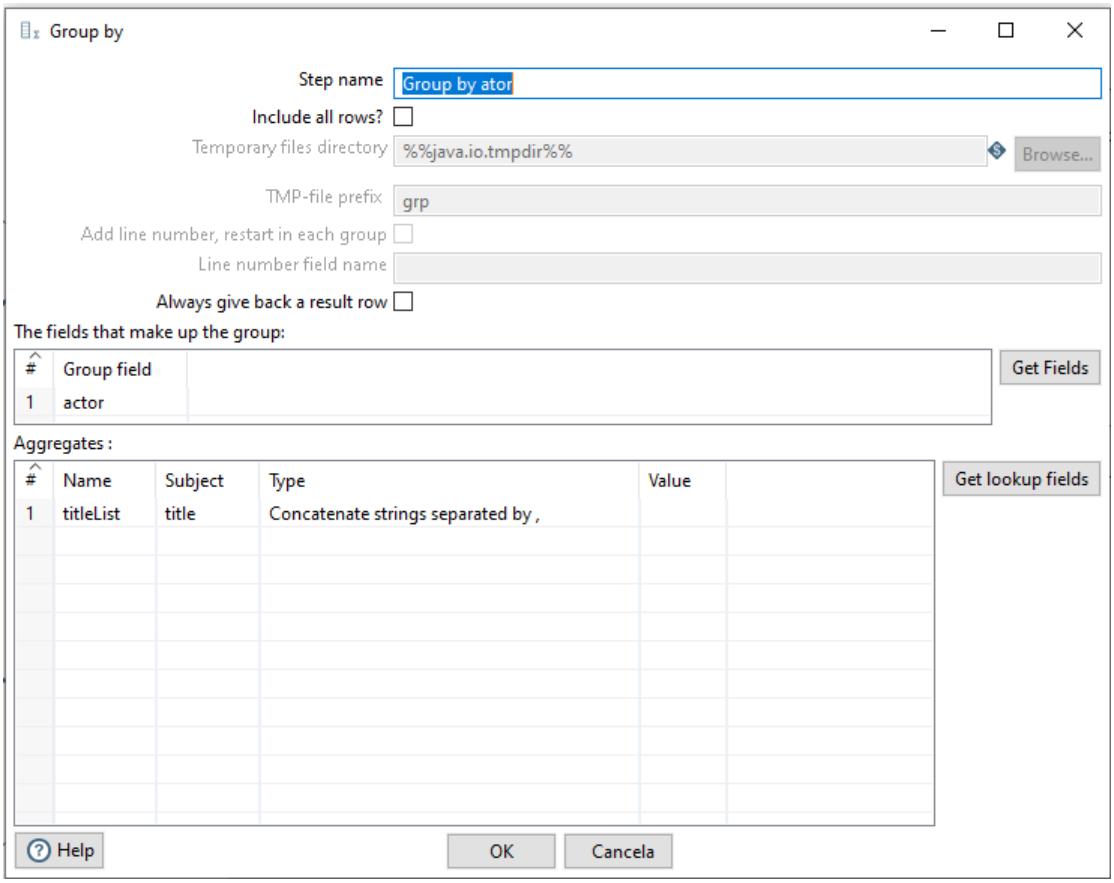

Figura 6 Configuração do Group by

- Lista de atores que já faleceram
  - A partir do campo "*date_of_death*" é feita um filtro de todos os atores que não têm este campo a "*null*".

Figura 7 Processo para os atores que ja faleceram.

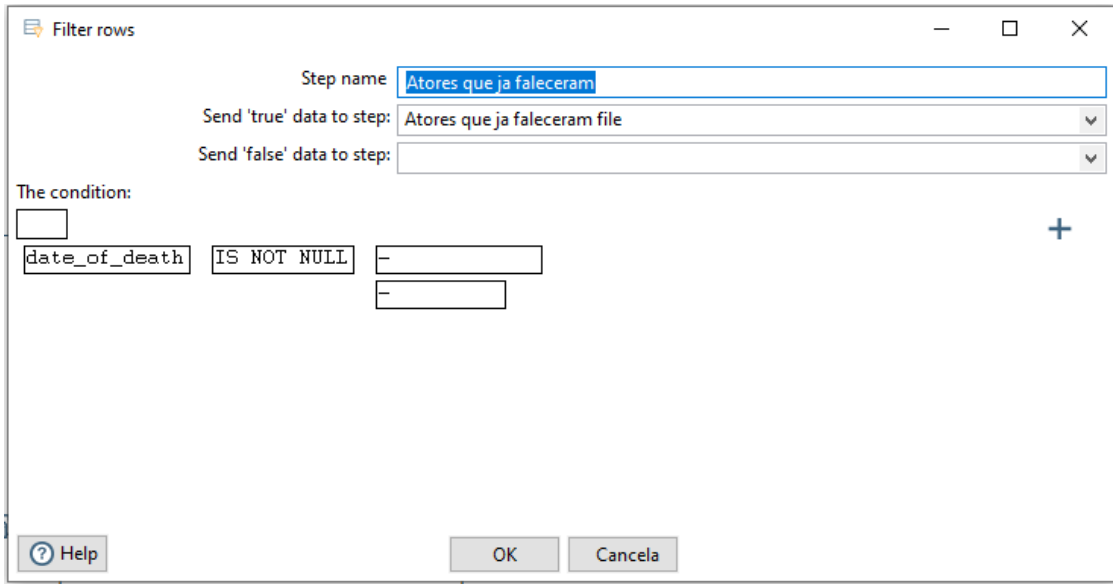

Figura 8 Configuração do filtro utilizado.

- Lista das longas metragens
    - Para este passo os dados foram filtrados a partir da sua duração, em minutos, tendo esta que ser maior ou igual que 70min. Após essa filtragem são removidos alguns dados não necessários e é feita uma ordenação pela duração do filme.

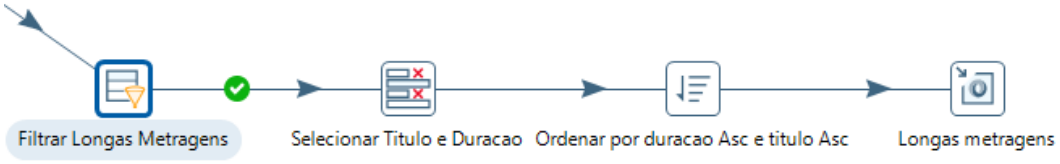

Figura 9 Processo completo para a lista de longas longas metragens

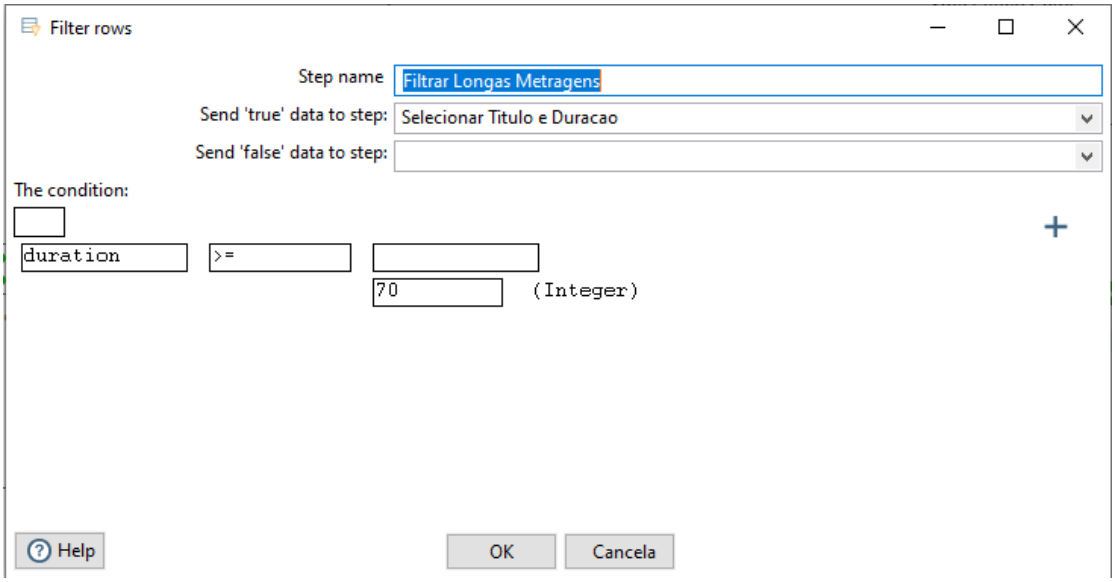

Figura 10 Configuração da filtragem da duração dos filmes

Para o processo de listagem dos filmes do *IMDB* disponíveis na *Netflix* são carregados os dados dos filmes da *Netflix* e utilizados os dados já carregados do *IMDB*

- Dados do *IMDB*
  - É feita uma remoção de dados ao arquivo de filmes do *IMDB* e uma ordenação pelo título dos filmes.

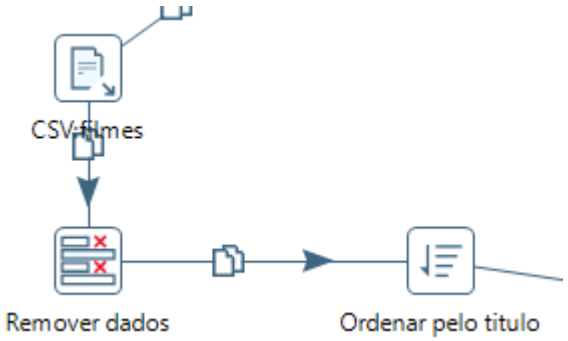

Figura 11 Processo de Remoção de dados e ordenação dos filmes.

- Dados Netflix
  - Nestes dados é feita uma remoção de colunas que não serão utilizadas e posteriormente é feita uma filtragem para se obter só os filmes e uma ordenação destes filmes por ordem alfabética.

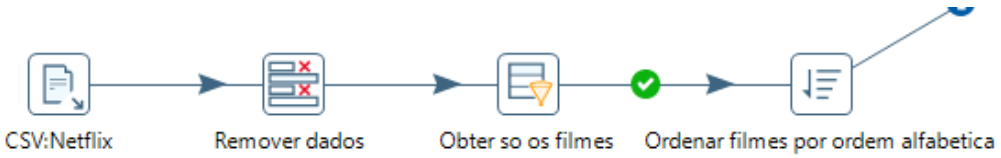

Figura 12 Processo completo do tratamento dos dados

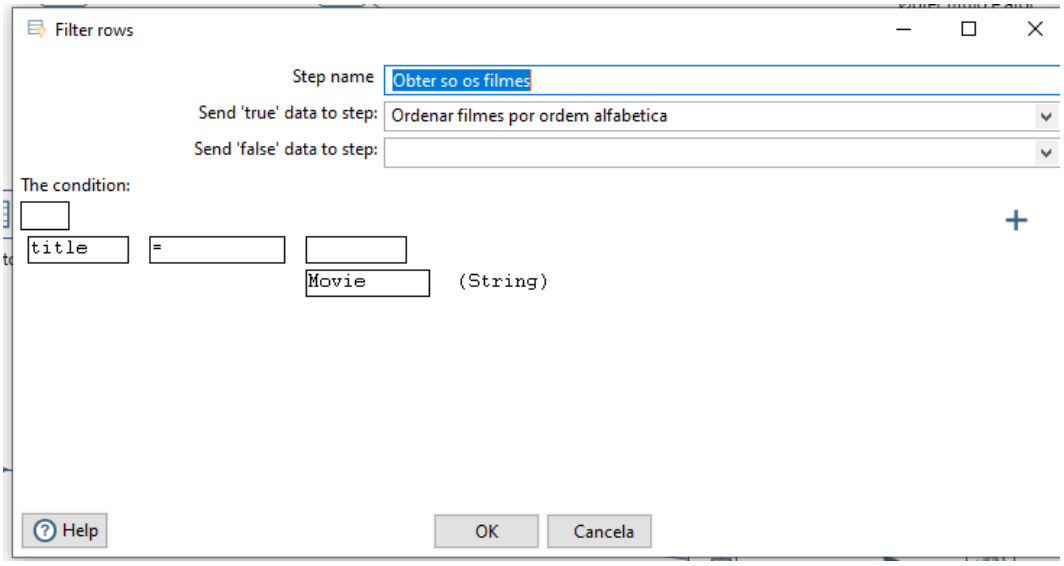

Figura 13 Configuração do filtro para obter só os filmes

- Para obter a listagem dos filmes quês estão disponíveis na *Netflix* foi utilizado o operador "*Merge rows (diff)*" para conseguir unir os dados dos dois arquivos. Na configuração deste operador o campo "flag fieldname" foi definido como "identical". Desta forma é possível só extrair os dados que são idênticos a partir da key selecionada. Sendo que a key escolhida foi o titulo dos filmes.

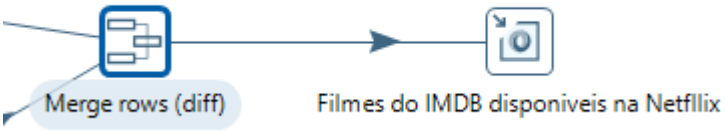

Figura 14 Parte final do processo de extraçao dos filmes comuns.

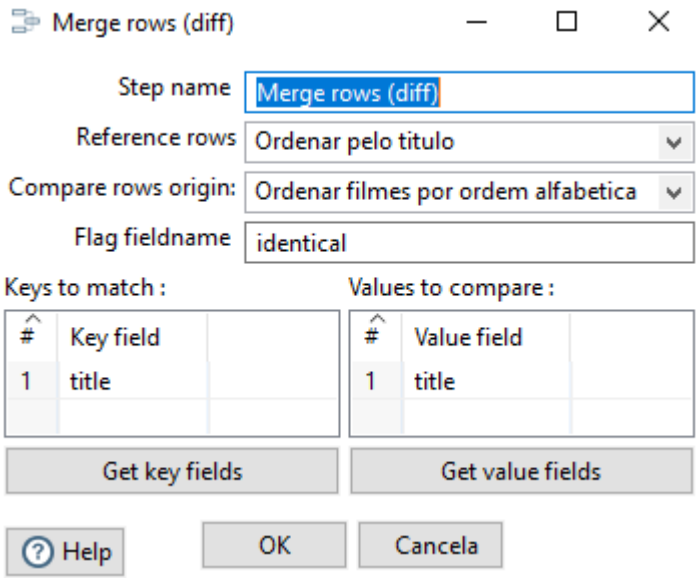

Figura 15 Configuração do Merge rows (diff)

Resultados do projeto

- Listagem de todos os filmes de cada ator, sendo que a listagem está ordenada alfabeticamente pelo nome do ator;
- Lista de atores que já faleceram;
- Lista de longas metragens existentes;
- Filmes do *IMDB* disponíveis na *Netflix*;

Dificuldades

- Os ficheiros utilizados serem de grandes dimensões o que consumia muitos recursos da máquina fazendo com que esta bloqueasse por alguns minutos;
- Falta de memória na máquina para conseguir efetuar as transformações em separado.
- Devido a alguma falta de criatividade/imaginação não surgiram mais ideias para elaborar mais o trabalho.
- Os arquivos que foram exportados são todos em formato json para consumir menos recursos.

Propostas de melhoria

- Os dados dos filmes e dos atores poderiam ser inseridos numa base de dados de forma a consumir menos recursos da máquina.

Conclusão

Com a realização deste trabalho foi possível consolidar os conhecimentos obtidos durante as aulas, melhorar a forma de pesquisa de informação quando surgiram dúvidas e principalmente entender os processos ETL e a utilidades destes.