INSTITUTO POLITÉCNICO
DO CÁVADO E DO AVE
ESCOLA SUPERIOR DE TECNOLOGIA

RELATÓRIO DE TRABALHO PRÁTICO

Relatório do Trabalho Pratico ISI

LUÍS MARTINS

ALUNO Nº 16980

Trabalho realizado sob a orientação de:
Luís Ferreira

Integração de Sistemas de Informação

Licenciatura em Engenharia de Sistemas Informáticos

Barcelos, Novembro de 2020

Índice

Lista de Figuras

Introdução

Neste relatório viso descrever o processo completo da realização do primeiro projeto da unidade curricular Integração de Sistemas de Informação. Este primeiro trabalho tem como objetivo a aplicação e experimentação de ferramentas em processos de ETL (Extract, Transformation and Load), inerentes a processos de Integração de Sistemas de informação ao nível dos dados.

Pretende-se que sejam desenvolvidos processos de ETL que envolvam *scripts* próprias ou que recorram a ferramentas sugeridas pelo professor como o Pentaho Kettle, Microsoft SQL Server Integration Services (MSSIS), Knime, Talend open studio, ou outras.

O que é o ETL?

ETL é um tipo de data integration em três etapas (extração, transformação, carregamento) usado para combinar dados de diversas fontes. Ele é comumente utilizado para construir um data warehouse. Nesse processo, os dados são retirados (extraídos) de um sistema-fonte, convertidos (transformados) em um formato que possa ser analisado, e armazenados (carregados) em um armazém ou outro sistema. Extração, carregamento, transformação (ELT) é uma abordagem alternativa, embora relacionada, projetada para jogar o processamento para o banco de dados, de modo a aprimorar a performance.

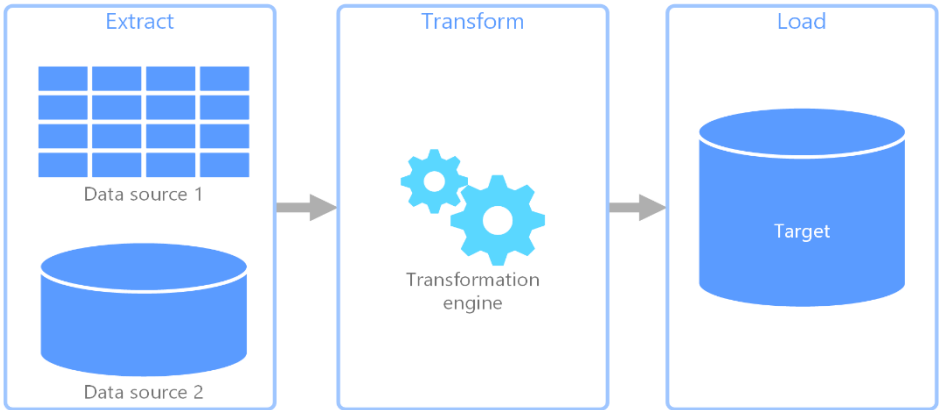

Figura 1 - Esquema do Processo ETL

Descrição do Problema

Com a evolução das plataformas de *streaming* online (Netflix, Hulu, etc.) a demanda de serviços de Tv tradicionais começam a baixar a cada dia que passa, vários estudos efetuados pelas próprias empresas que fornecem tais serviços concluem que o número de pessoas que vê televisão baixa cada dia que passa.

O objetivo deste trabalho é, utilizando uma dataset da Netflix fornecida por um site, trabalhar com a ferramenta *Pentaho Kettle* para filtrar os dados para obtermos o resultado pretendido.

Neste caso pretendemos separar as categorias de filmes acessíveis na plataforma Netflix, em filmes dirigidos a crianças e filmes dirigidos a adultos, isto é possível utilizando o Sistema de classificação de filmes (Neste caso utilizamos o sistema americano).

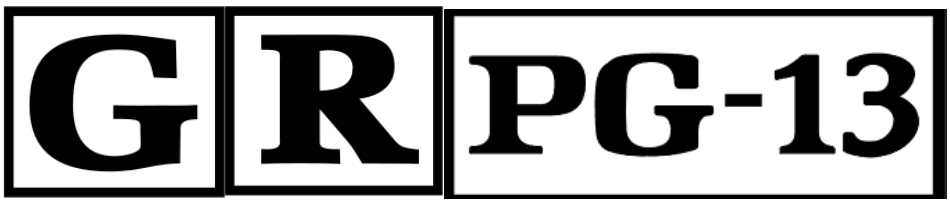

Figura 2 - Sistema de classificação de filmes

- G – Audiência geral, normalmente atribuído a filmes dirigidos para todas as idades
- R – Restrito menores de 17 anos exigem o acompanhamento dos pais ou responsável adulto (Semelhante a NC-17 onde ninguém abaixo dos 17 anos pode assistir)
- PG - sugestão de orientação parental - Algum material pode não ser adequado para crianças (Variação PG-13 onde orientação paternal é fortemente aconselhada)

Desenvolvimento

Transformação

Utilizando a ferramenta Pentaho Kettle criamos uma transformação destinada a converter e filtrar dados de um ficheiro "*.CSV*" para um ficheiro XML como também para vários ficheiros JSON.

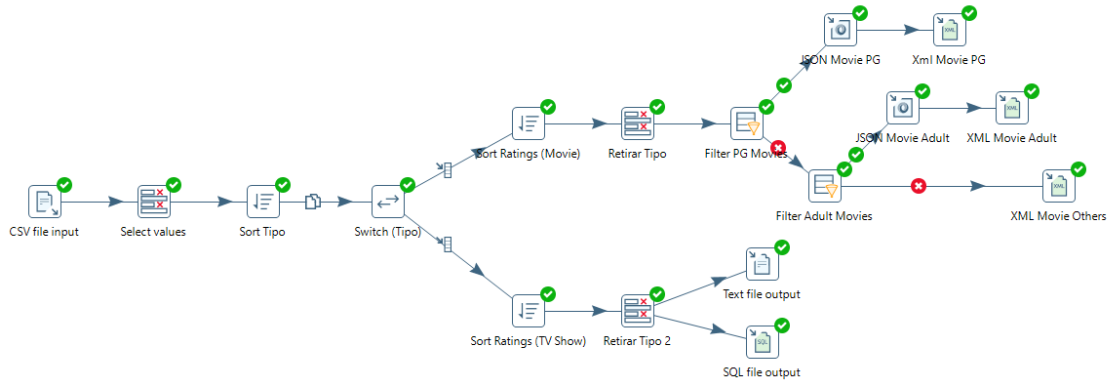

Figura 3- Transformação completa

Primeiro começamos por aceitar o ficheiro CSV como input e eliminamos certas áreas não uteis para este trabalho, depois ordenamos os dados por tipo (Movies, TV Shows).

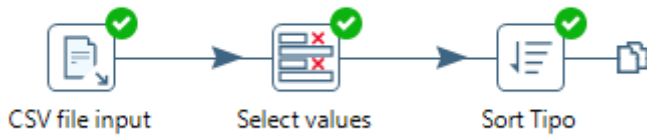

Figura 4 - Primeiro passo (Transformação)

Depois usamos um *switch / case* para separar os dois tipos ordenados anteriormente para nos focarmos nos filmes.

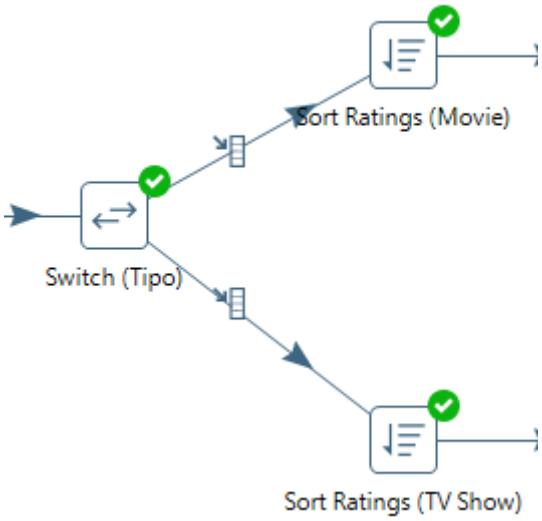

Figura 5 - Switch/Case

De seguida, se seguimos o ramo dos filmes voltamos a efetuar partes do primeiro passo onde desta vez retiramos a coluna do tipo como também ordenamos os filmes por "*ratings*", ou por as a classificações que vimos anteriormente, para depois filtrarmos as mesmas em ramos diferentes.

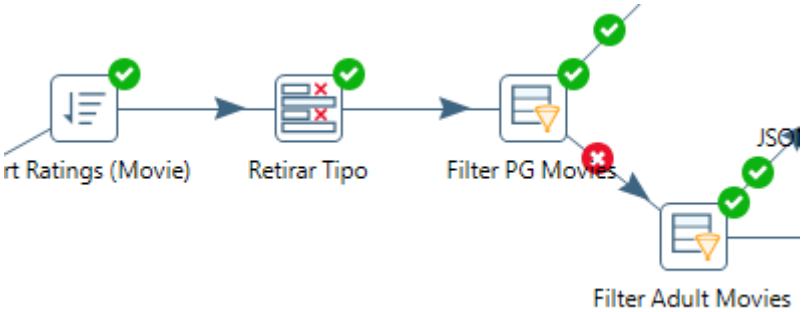

Figura 6 - Ordenação de ratings e filtros

No próximo passo, utilizando expressões regulares, usamos a ferramenta "*Filter rows*" para separar as várias classificações dos filmes ….

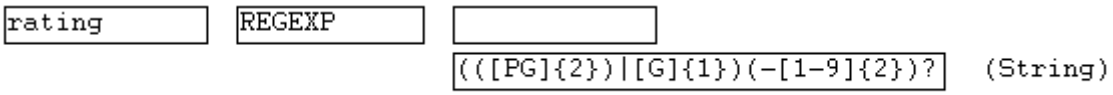

Figura 7 – Expressão regular para filtrar Filmes PG

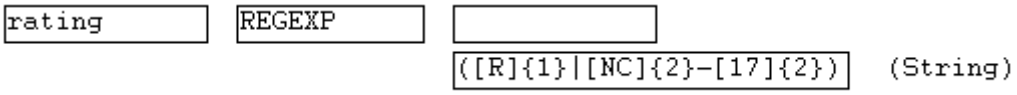

Figura 8 - Expressão regular para filtrar Filmes Adultos

Por último, convertemos os dados filtrados para ficheiros JSON e para um ficheiro XML para serem usados mais à frente.

Como algo extra decidi utilizar os TV Shows que foram separados da coluna dos "tipos" para criar uma integração a uma base de dados SQL, neste trabalho decidi usar o *"PostgreSQL"* visto que foi o gerador de base de dados utilizado na unidade curricular de Base de Dados no ano letivo anterior .

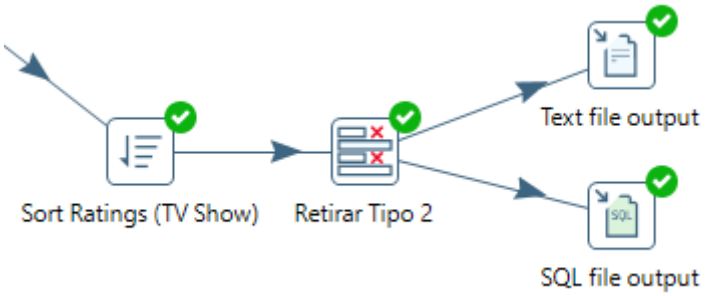

Figura 9 - Ramo TV Shows

Para efetuar a integração SQL temos de criar uma conexão nova que se conecte à base de dados criada por mim no servidor PostgreSQL para este trabalho.

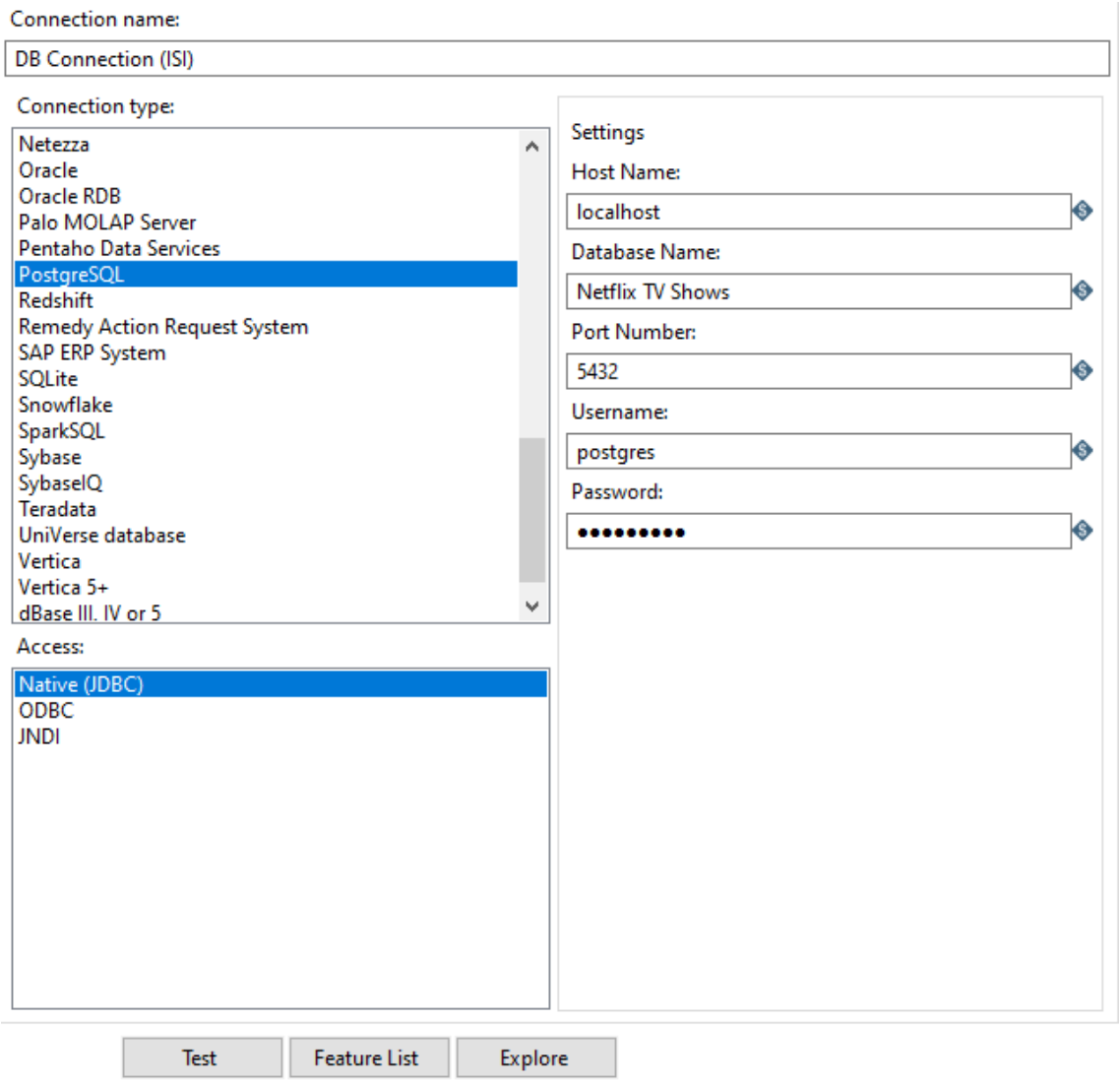

Figura 10 - Conexão à base de dados

Fazendo um pequeno teste vemos que a conexão foi um sucesso.

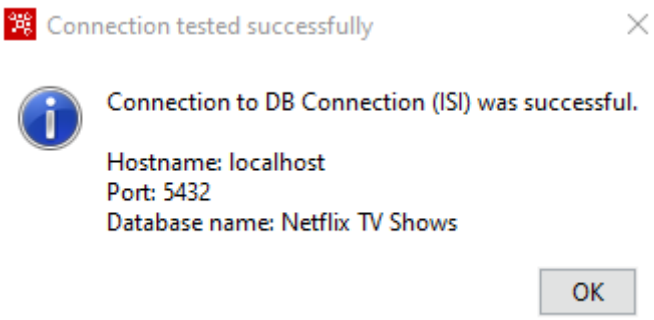

Figura 11 - Teste de conexão

De seguida usando a node "SQL Output" configuramos de modo a criar uma table, usando um querie que acomode todos os dados processados até agora.

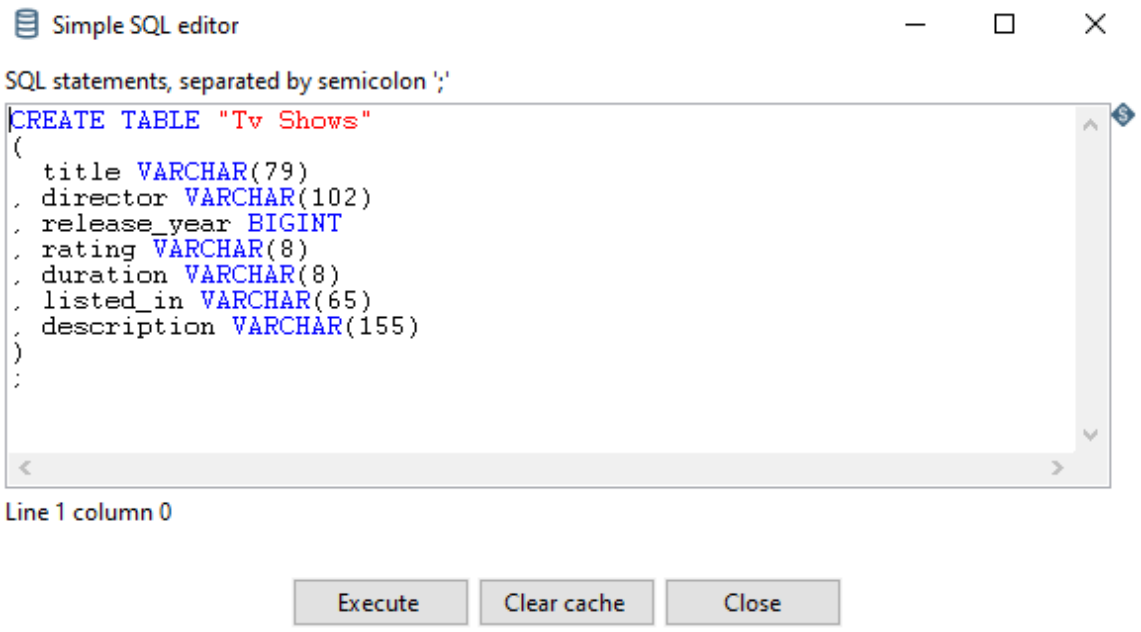

Figura 12 - Querie que cria table TV Shows

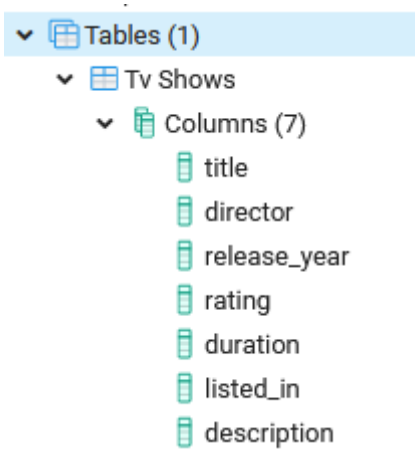

Figura 13 - Table Criada na BD

Com isto está concluída a transformação criada para este processo funcionar…

Job

Jobs no Pentaho Kettle são utilizados para coordenar certas atividades ETL como a ordem de transformações, preparações de execução como verificação da existência de ficheiros etc.

Para este trabalho utilizei um Job para criar as pastas necessárias para o trabalho, como também para transformar os ficheiros XML obtidos a partir da transformação, em ficheiros HTML como também enviamos por email para um destinatário da nossa escolha (Neste caso enviei para mim próprio).

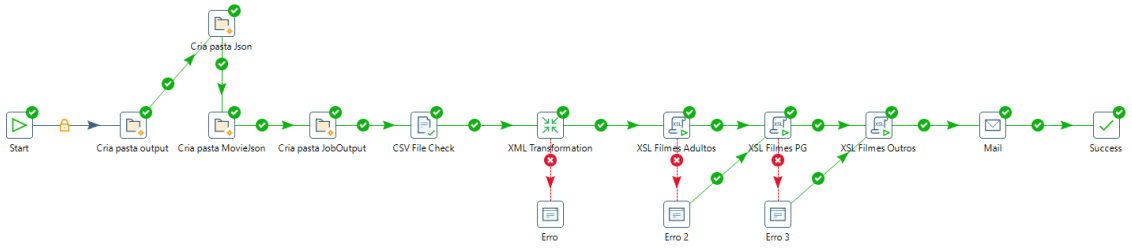

Figura 14 -Job criado para o trabalho

Depois de criar as pastas e verificar se o ficheiro que contem os dados existe, invocamos a transformação e, depois, procedemos com a transformação dos ficheiros XML.

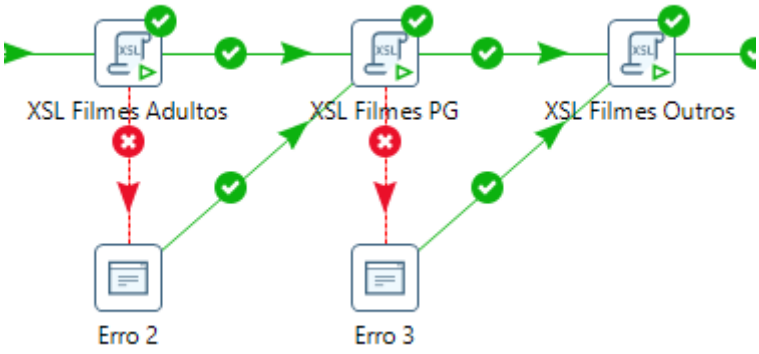

Figura 15 - Transformação XML usando XSL

No final obtemos três ficheiros HTML com a informação processada e como também com *hyperlinks* que levam a cada ficheiro.

Conclusão

Com a conclusão deste trabalho acredito que os meus conhecimentos em relação à área de integração de sistemas de informação foram expandidos, e sei agora a importância desta área no mundo profissional.

Embora não tenha explorado tudo que o programa tenha a dar como, explorar o acesso a serviços Web remotos, acredito que tenho um conhecimento básico do potencial desta ferramenta.

Bibliografia

- Material disponibilizado no Moodle
- https://help.pentaho.com/
- https://stackoverflow.com/
- https://www.wikipedia.org/