

**Instituto Politécnico do Cávado e do Ave**

Integração de Sistemas de Informação

**Extract, Transform & Load – IMDB Movie Dataset**

Engenharia de Sistemas Informáticos

Duarte Ribeiro de Melo – 21149

Docente Óscar Ribeiro

Barcelos, Portugal

15 novembro de 2022

# Resumo e estrutura do documento

Com o avanço das tecnologias é de extrema importância o cuidado na manipulação, transformação, armazenamento e outras operações realizadas sobre dados, assim como a adaptação dos mesmos a novas tecnologias e, acima de tudo, garantir a flexibilidade e possibilidade de integrar os mesmos em soluções diferentes.

Serve então o presente relatório para demonstrar o trabalho desenvolvido pelo aluno Duarte Ribeiro de Melo, relativo ao Trabalho Prático 01 da Unidade Curricular de Integração de Sistemas de Informação, que toca nos temas mencionados supra.

Numa primeira fase do documento consta uma introdução ao tema, objetivos do projeto, demonstração do *dataset* escolhido e das ferramentas escolhidas pelo aluno.

Numa segunda fase está presente a explicação da solução mais detalhada, tanto a parte realizada no programa *Pentaho Kettle*, em *Python* ou em *AngularJS*.

Por fim, alguns tópicos que visam dar como concluído o projeto e algumas considerações finais.

Índice

[Resumo e estrutura do documento 2](#_Toc119373614)

[Índice de Figuras 4](#_Toc119373615)

[Introdução 5](#_Toc119373616)

[Objetivos 6](#_Toc119373617)

[Dados utilizados no processo ETL 7](#_Toc119373618)

# Índice de Figuras

[Figura 1 - Exemplo de registos dataset title.basics.tsv.gz 8](#_Toc119373575)

[Figura 2 - Exemplo de registos dataset title.ratings.tsv.gz 8](#_Toc119373576)

# Introdução

Nos dias que correm, os nossos dados estão presentes em sistemas informáticos relacionados com as áreas mais sensíveis e importantes da nossa vida, como a saúde, educação, segurança, entre outros. É imprescindível, mas exigente, garantir que estes dados se mantêm coesos, seguros, atualizados e passíveis de serem utilizados em sistemas em constante atualização, que sofrem reestruturações e reformulações com uma rapidez nunca antes vista.

Sistemas informáticos da área da saúde, banca, entre outras que se prendiam por sistemas antigos pela sua segurança e dificuldade de atualização dos mesmos, veem agora a necessidade de se adaptar a um ecossistema totalmente diferente, onde os dados devem estar acessíveis num smartphone, numa caixa de multibanco, num computador, num tablet ou num terminal de pagamento, permitindo que sejam realizadas operações entre estes diferentes sistemas mencionados.

Resumindo, é um grande tópico da área informática a segurança, coesão e flexibilidade dos dados, assim como a grande necessidade de disponibilização dos mesmos em formatos transversais a diversas soluções.

Surge então o tema da integração dos dados, abordado nesta Unidade Curricular, ao qual pertencem os processos de ETL – *Extract, Transform and Load* – extração, transformação e carregamento.

Com base nisto, optou-se pela utilização do *software* Pentaho Kettle, no qual se desenvolveu a abordagem às três letras do ETL, e ainda a utilização de outras ferramentas para a análise e visualização dos resultados obtidos.

# Objetivos

O objetivo principal deste projeto é a extração e transformação de dados provenientes de um *dataset*, nomeadamente o da *Internet Movie Database* (IMDB), e consequente carregamento dos dados noutros sistemas, neste caso, numa base de dados IMDB – todo este processo é realizado no Pentaho Kettle. Após este carregamento, deve ser possível a visualização dos dados num *dashboard* desenvolvido em *AngularJS*, que consegue obter os dados presentes na base de dados via web API desenvolvida em *Flask*, framework de *Python*.

Foram escolhidos dois ficheiros presentes no *dataset*, nomeadamente o *title.basics.tsv.gz* e o *title.ratings.tsv.gz* , que foram agrupados e tratados na ferramenta de ETL. O resultado do agrupamento e tratamento dos dados destes dois documentos é então exportado para a base de dados MongoDB, para documentos XML e para um ficheiro JSON – separado por tipo de título (filme, curta, episódios, etc.).

# Dados utilizados no processo ETL

Para que a realização deste trabalho fosse possível era necessário escolher um *dataset* disponível na Internet, de preferência com alguma complexidade, de forma a permitir algumas operações sobre o mesmo. Assim sendo, foi escolhido o *dataset* disponibilizado pelo IMDB no seguinte URL - <https://datasets.imdbws.com/> - este tem mais de 10 milhões de registos em alguns ficheiros e diversos campos relativos a produções audiovisuais.

Destes dados disponibilizados pelo IMDB, foram apenas utilizados dois ficheiros - o *title.basics.tsv.gz* e o *title.ratings.tsv.gz* – o primeiro contém informação geral dos *titles* (*titles* incluem qualquer tipo de produção audiovisual – filmes, séries, episódios, vídeos, etc. – presente no IMDB) e o segundo contém informação sobre as avaliações atribuídas pelos utilizadores da plataforma IMDB aos *titles* visualizados por estes.

É também importante mencionar que a descompactação destes dados é realizada pelo próprio *Kettle*, dado que a extração é feita pelo mesmo, diretamente do URL acima.

Passando à explicação dos campos presentes nos ficheiros obtidos do *dataset*, no *title.basics.tsv.gz*:

* tconst – identificador único de cada *title* (produção audiovisual presente no IMDB)
* titleType – o tipo de produção (filme, curta metragem, série, etc.)
* primaryTitle – o título mais usado/comum para a produção
* originalTitle – o título original, na linguagem original, da produção
* isAdult – indica se o filme é para adultos (valor 1) ou não (valor 0)
* startYear – ano de lançamento, no caso de ser uma série, o ano em que começou
* endYear – só se aplica nas séries e é o ano de término, tem o valor de ‘\N’ para outro tipo de produções
* runtimeMinutes – duração da produção em minutos
* géneros – até três géneros associados à produção (Drama, Biografia, etc.)

No *title.ratings.tsv.gz*:

* tconst – identificador único de cada *title* (produção audiovisual) ao qual se refere este *rating*/avaliação
* averageRating – valor médio das avaliações
* numVotes – número de votos por parte dos utilizadores em relação a esta produção

É importante mencionar que no caso em que não exista informação para algum destes campos, é encontrado um ‘\N’ no mesmo.

Exemplo de registos destas tabelas:

*Uma imagem com texto

Descrição gerada automaticamente*

Figura 1 - Exemplo de registos dataset title.basics.tsv.gz

Uma imagem com texto

Descrição gerada automaticamente

Figura 2 - Exemplo de registos dataset title.ratings.tsv.gz

# Ferramentas utilizadas

Neste tópico serão abordadas as ferramentas utilizadas para o desenvolvimento do projeto, assim como uma breve descrição das mesmas e em que sentido foram necessárias para desenvolver este trabalho.

## Pentaho Data Integration – Kettle

O Pentaho Data Integration (PDI) providenciou ferramentas para a realização dos processos ETL – *Extract, Transform and Load*. Estas facilitaram o processo de extração, correção e armazenamento dos dados utilizando um sistema de simples compreensão, sem necessidade de escrever código e de rápida implementação.

Para que este programa corra, é necessário ter o Java instalado no computador, assim como definir a variável do sistema JAVA\_HOME.

No caso específico das transformações e *jobs* desenvolvidos neste trabalho, dado o grande volume da dados, foi necessário expandir a memória RAM máxima utilizada pelo programa para 8GB:

Uma imagem com texto

Descrição gerada automaticamente

Figura 3 - Alteração memória máxima usada pelo Pentaho Kettle

## Python com framework Flask

Ferramentas utilizadas

Solução (flow ETL)

(explicar fases, ver Git e Notion)

Solução em Pentaho Kettle

Solução em Python Flask (API)

Solução em AngularJS

Conclusão

Lições aprendidas

Trabalho futuro

Apreciação final

Webgrafia