Relatório do trabalho da disciplina de Engenharia de Software

ETL (Extract, Transformation and Load)

Luís Brito - 14076

Engenharia de Sistemas Informáticos

11/2020

ESCOLA SUPERIOR DE TECNOLOGIA IPCA

Índice

Conteúdo

Índice de figuras

Introdução

Com este trabalho da Disciplina de Integração de Sistemas de Informação (ISI) pretende-se focar 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 disponíveis como o Pentaho Kettle, Microsoft SQL Server Integration Services (MSSIS), Knime, Talend open studio, ou outras.

Neste trabalho foi também usado Kaggle como base de dados. Kaggle é uma comunidade online de cientistas de dados e profissionais de "machine learning". O Kaggle permite que os usuários encontrem e publiquem conjuntos de dados, explorem e criem modelos em um ambiente de ciência de dados baseado na web, trabalhem com outros cientistas de dados e engenheiros de "machine learning "e participem em competições para resolver desafios de ciência de dados.

Introdução ao problema

YouTube (o site de compartilhamento de vídeo mundialmente famoso) mantém uma lista dos vídeos mais populares na plataforma. De acordo com a revista Variety, “para determinar os vídeos mais populares do ano, o YouTube usa uma combinação de fatores, incluindo a medição das interações dos usuários (número de visualizações, compartilhamentos). No entanto eles não são os vídeos mais vistos em geral no ano inteiro”. Os melhores desempenhos na lista de tendências do YouTube são os videoclipes (como o famoso e viral “Gangnam Style”), performances de celebridades e / ou reality shows na TV e os vídeos virais aleatórios pelos quais o YouTube é conhecido.

O conjunto de dados usado é um registo diário dos vídeos mais populares do YouTube.

Neste caso o problema a resolver é apresentar os vídeos mais populares em Portugal de acordo com as visualizações de forma descendente.

Para resolver o problema foi usado uma ferramenta para carregar e exportar dados com o nome de PDI (Pentaho Data Integration). A partir desta ferramenta é possível fazer uma “transformation” e um “Job” de modo aos dados ficarem corretamente tratados para serem apresentados ao utilizador.

Importação de dados

Para a realização deste trabalho foi usado como base dados um ficheiro de formato CSV (valores separados por vírgula, este contém bastante campos, no entanto foram removidos alguns de acordo a ter um ficheiro mais compacto apenas com o essencial a trabalhar.

video_id	trending_	title	category_id	views
n1WpP7ic	17.14.11	Eminem - Walk On Water (Audio) ft. Beyonc	10	17158579
0dBIkQ4M	17.14.11	PLUSH - Bad Unboxing Fan Mail	23	1014651
5qpjK5Dg	17.14.11	Racist Superman | Rudy Mancuso, King Bach	23	3191434
d380meD	17.14.11	I Dare You: GOING BALD!?	24	2095828
2Vv-BfVo	17.14.11	Ed Sheeran - Perfect (Official Music Video)	10	33523622

Figura 1 Campos do ficheiro CSV

Como apresenta a imagem acima podemos ver que a base de dados é constituída por os seguintes campos:

- Video_ID ;
- Trending_Date ;
- Title ;
- Category_ID ;
- Views .

Tratamento de dados – Transformation

Para iniciar a "Transformation" foi usado o seguinte método para tratar os dados.

Figura 2 Esquema da Transformation

Inicialmente é indicado o path(caminho do ficheiro) para o ficheiro CSV que contém a base de dados. No entanto para tornar o mesmo mais fácil de trabalhar é usado uma variável interna do PDI que identifica onde o ficheiro se encontra no computador *${Internal.Entry.Current.Directory}.*

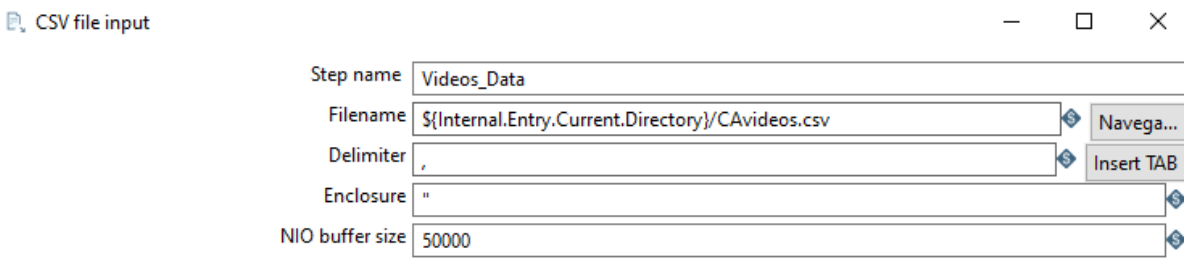

Figura 3 Inserir o caminho relativo

Após ser inserido o caminho do ficheiro é usado a opção *Obtem campos* com o objetivo de recolher da base de dados os campos necessários a trabalhar.

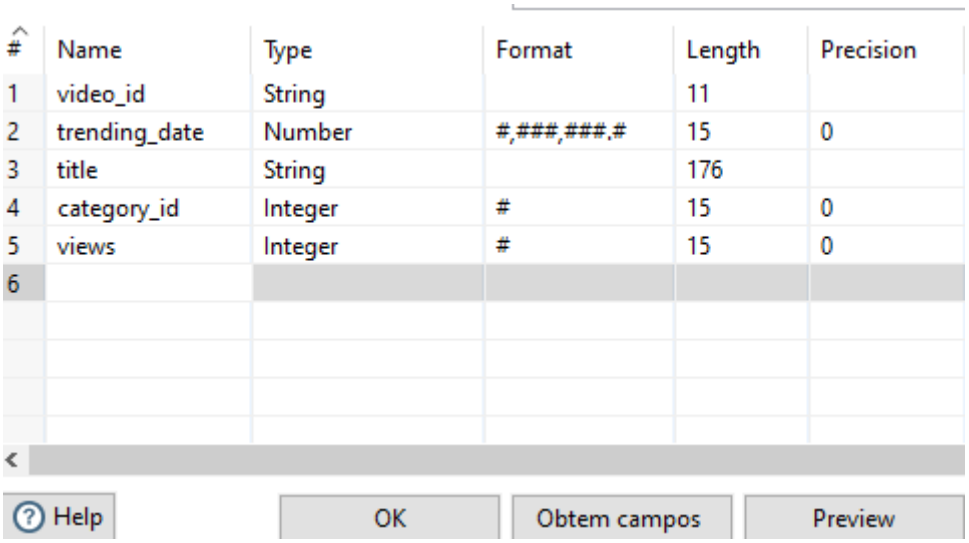

Figura 4 Recolha dos campos da base de dados

Também é possível caso seja desejado visualizar se os dados estão corretos a partir da opção *Preview* como indica a figura acima.

De seguida é realizada a ordenação do campo *Views* de forma descendente a partir da opção *Sort Rows* como indica a figura abaixo.

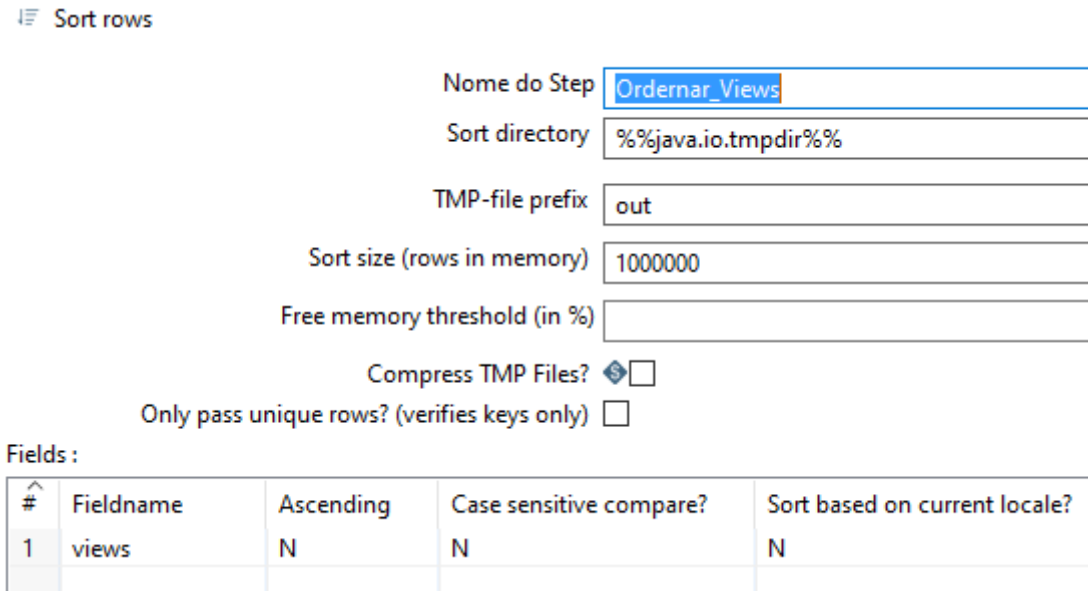

Figura 5 Ordenação do campo Views

De seguida é realizado o mesmo método para ordenar a data e para finalizar é usado o campo *XML Output* para colocar os dados em um ficheiro XML.

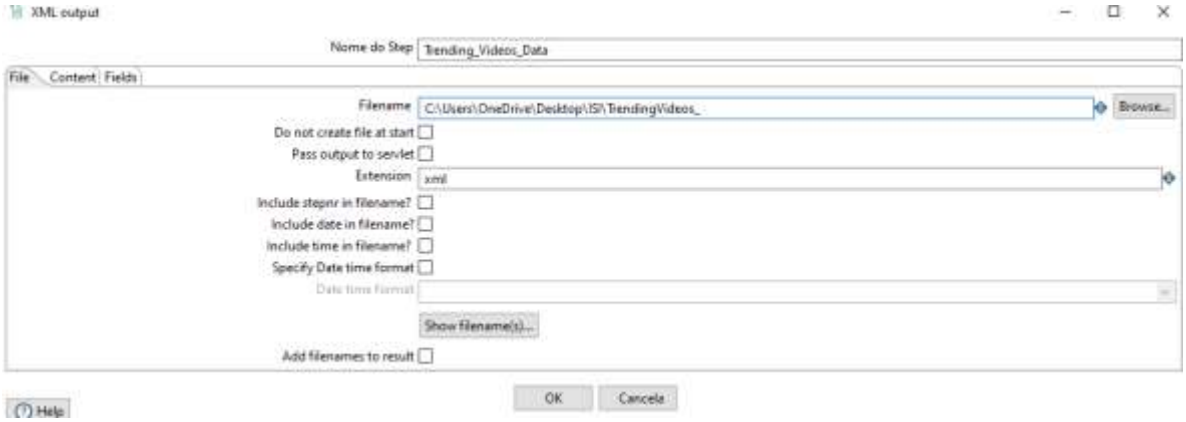

Figura 6 Output dos dados para XML

Organização dos Dados – Job

Para apresentar os dados ao utilizador final é necessário a aplicação de um *Job,* isto é a aplicação de uma folha de estilos XSL.

Isto para que os dados estejam em um formato de tabela de forma a ser mais fácil a visualização de dados.

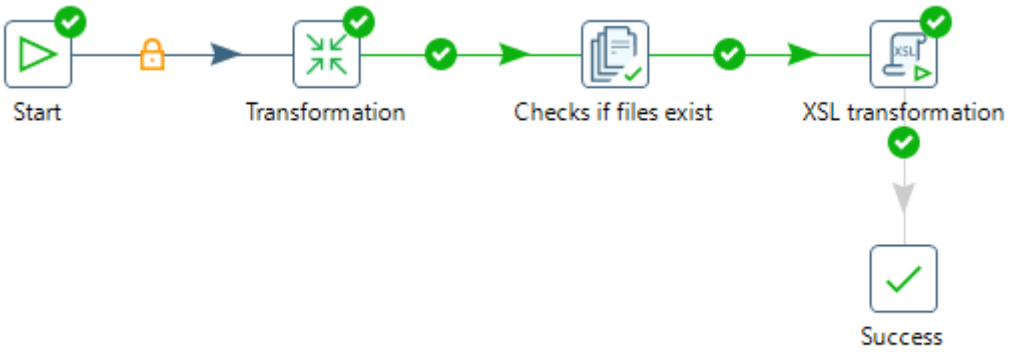

Figura 7 Esquema do Job

Este *Job* é iniciado por carregar a *Transformation mencionada anteriormente. Após esse passo estiver completo o programa irá verificar se o ficheiro XML existe e por fim se esse for o caso irá ocorrer a aplicação da folha de estilos XSL no ficheiro XML resultado da transformação.*

De modo a construir a tabela na folha de estilos XSL foi usado o seguinte método.

Inicialmente foi colocado um pequeno design á tabela tal como pode-se ver na figura abaixo

```
<style>
    table{
    width:40%;
    }
    table, th ,td{
    padding: 5px;
    text-align: left;
    }
</style>
```

Figura 8 Código do design da tabela ficheiro XSL

Após isto é construído o *Header* da tabela tal como adicionado um pouco de design á mesma.

```
<th bgcolor="#9acd32">
    Title
</th>
<th bgcolor="#9acd32">
    Views
</th>
```

Figura 9 Código do design Header tabela XSL

De seguida é usado o ciclo *for each* para selecionar em cada *Row* as variáveis *title* e *Views.*

Isto para que seja obtido por cada linha (Row) o título do vídeo tal como as suas visualizações como indica a linha 29 e 32.

```
26                  <xsl:for-each select="/Rows/Row">
27                  <tr>
28                      <td>
29                          <xsl:value-of select="title"/>
30                      </td>
31                      <td>
32                          <xsl:value-of select="views"/>
33                      </td>
34                  </tr>
35                   </xsl:for-each>
```

Com o *Job* completo é criado um ficheiro HTML para o utilizador final visualizar os dados.

Trending Videos Portugal

Title	Views
Sanju | Official Trailer | Ranbir Kapoor | Rajkumar Hirani | Releasing on 29th June	21739537
Cardi B, Bad Bunny & J Balvin - I Like It [Official Music Video]	20723565
Pusha T The Story Of Adidon (Drake Diss) (WSHH Exclusive - Official Audio)	8654006
Drake - I'm Upset	7766948
Kaala (Tamil) - Official Trailer | Rajinikanth | Pa Ranjith | Dhanush | Santhosh Narayanan	7147686

Figura 10 Visualização dos dados HTML

Referências

https://www.kaggle.com/