

RELATÓRIO DE TRABALHO PRÁTICO

Trabalho Prático nº1

DIOGO ROCHA

ALUNO Nº 16966

Trabalho realizado sob a orientação de:
Luís Ferreira

Integração de Sistemas Informáticos

Licenciatura em Engenharia de Sistemas Informáticos

Índice

Conteúdo

TRANSFORMAÇÃO	3
Ficheiro CSV	4
Formatação do ficheiro	4
Switch/ Case	7
Outputs	8
Carregamento para a base de dados	8
FICHEIRO XML	9
Ficheiro DTD	10
Output de XSL para HTML	11
XSL Transformation	12
HTML	12
Email	13
Output Email	14
BIBLIOGRAFIA	15

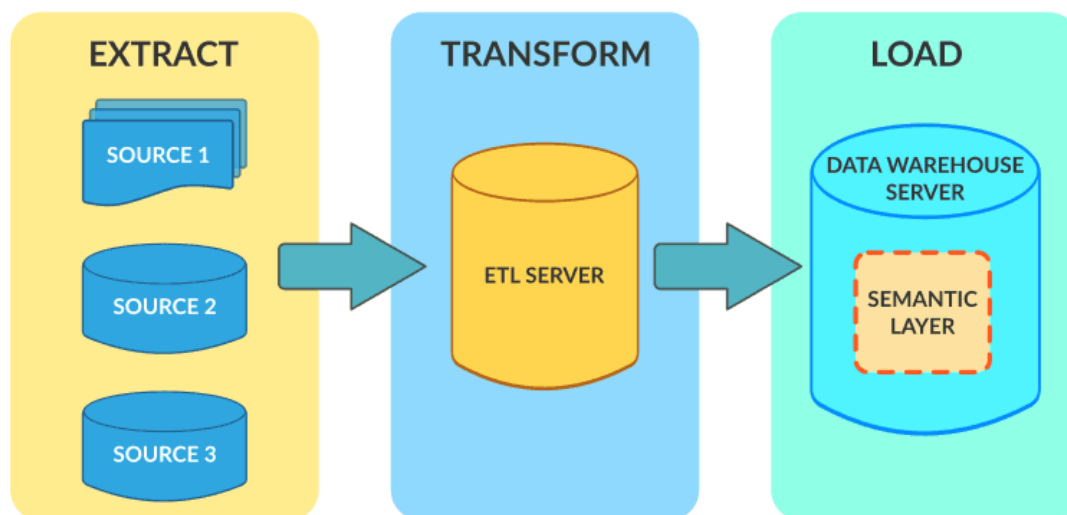
Índice de figuras

Figura 1- Transformação	3
Figura 2 - Ficheiro CSV	4
Figura 3 - Select Values.....	4
Figura 4 - Unique Rows.....	5
Figura 5 - Reple Nulls with ND	5
Figura 6 - String Operations	6
Figura 7 - Sort Rows	6
Figura 8 - Switch/ case.....	7
Figura 9 - Filter Rows.....	7
Figura 10 - Outputs	8
Figura 11 - Database Connection	8
Figura 12 - Insert / update	9
Figura 13 - XML	9
Figura 14 – Job	10
Figura 15 - DTD file.....	10
Figura 16- XSL.....	11
Figura 17- XSL transformation	12
Figura 18- HTML.....	12
Figura 19- Mail	13
Figura 20 - Output Email.....	14

Introdução

Este trabalho foi realizado no âmbito da disciplina de Integração de Sistemas de Informação, lecionada pelo professor Luís Ferreira. Foi nos proposto o uso de uma ferramenta ETL, para a qual optei por utilizar o Pentaho Kettle.

Com este trabalho 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.



Resumo

Comecei inicialmente por procurar um ficheiro .csv que continha vária informação sobre jogadores da Premier League da época 2018/2019.

Após aceder aos dados do ficheiro recorri a uma transformação em que inicialmente foi necessário filtrar alguns dados não relevantes e organizá-los de forma a que qualquer utilizador consiga perceber o contexto e trabalhar sobre os mesmos facilmente. Após ter filtrado estes mesmos dados, separei os ficheiros conforme as posições dos jogadores, de modo a ser mais fácil distingui-los, e por sua vez acabei por criar um ficheiro XML como output da transformação.

Posteriormente realizei um job, que verifica se o ficheiro csv existe, se existir então ocorrerá a transformação anteriormente, que será validada por um ficheiro DTD, caso seja validado então existe a transformação de XML para XSL e por fim é enviado um email ao destinatário sobre possíveis erros que poderão ter ocorrido durante o processo, se o job funcionar então é gerado o html com os jogadores da liga inglesa.

Desenvolvimento

Transformação

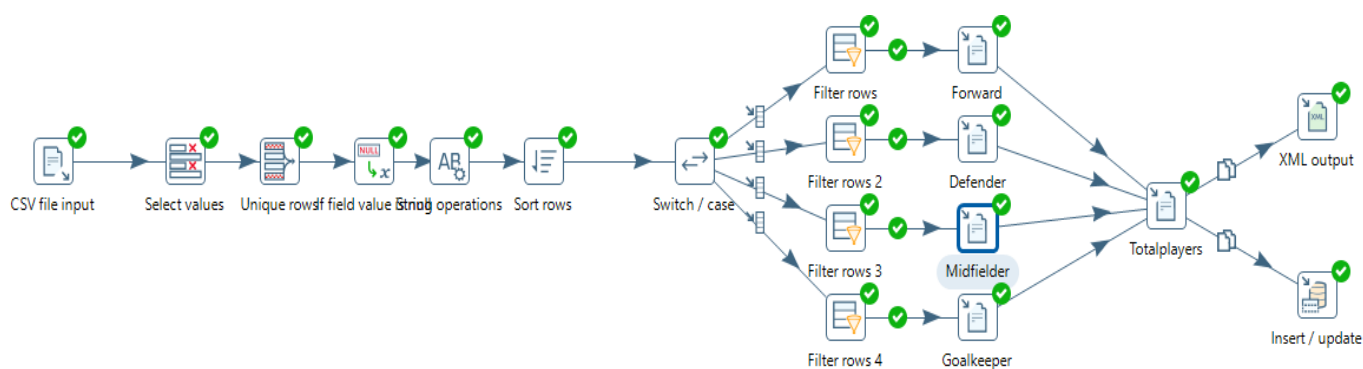


Figura 1- Transformação

Como foi dito acima, a transformação começa por pegar no ficheiro .csv no entanto como este ficheiro está completamente cheio de informação desnecessária procedi então à filtragem do mesmo.

Ficheiro CSV

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U																									
1	full_name	age	birthday	league	season	position	Current_Club	minutes_played_overall	minutes_played_home	minutes_played_away	nationality	appearances_overall	appearances_home	appearances_away	goals_overall																															
2	Aaron Cresswell	30	629683200	Premier League	2018/2019	Defender	West Ham United	1589,888	701	England	20	10	8	0	0	1	1	0	0	3	2	12,12	10	1	0	0	0	0	0	0	1.25	72	79	1589	1589	0.06	290	191	80	20						
3	Aaron Lennon	33	545526000	Premier League	2018/2019	Midfielder	Burnley	1217,487	730	England	16	5	9	1	1	0	1	0	0	4	2	2,20	8	12	1	0	0	15	0.07	0.07	0.18	0	1217	1.48	61	76	1217	1217	0.07	199	187	-1	10			
4	Aaron Mooy	30	653353200	Premier League	2018/2019	Midfielder	Huddersfield Town	2327,1190	1137	Australia	29	13	12	3	1	2	1	0	1	1	0	4	3	1	46	20	26	4	0	0	15	0.04	0.12	0.08	0.16	776	1.78	51	80	582	2327	0.15	147	233	-1	3
5	Aaron Ramsey	29	662169600	Premier League	2018/2019	Midfielder	Arsenal	1327,689	638	Wales	28	8	6	4	2	6	5	1	0	0	7	6	1	12	2	10	0	0	0	68	0.41	0.27	0.26	0.28	332	0.81	111	47	0	221	0	69	8	-1	5	
6	Aaron Rowe	20	968281200	Premier League	2018/2019	Forward	Huddersfield Town	69	14	55	England	2	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	3	69	35	0	0	-1	-1	-1	31						
7	Aaron Wan-Bissaka	22	880502400	Premier League	2018/2019	Midfielder	Crystal Palace	3135	1605	1530	England	35	18	17	0	0	3	1	2	0	0	12	7	5	41	17	24	5	1	0	09	0.09	0.09	0	0	1	18	76	90	523	1045	0.17	312	160	-1	22
8	Abdelhamid Sabiri	23	849139200	Premier League	2018/2019	Midfielder	Huddersfield Town	49	0	49	Morocco	2	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	5	51	16	25	0	0	-1	-1	-1	22						
9	Abdoulaye Doucouré	27	725846400	Premier League	2018/2019	Midfielder	Watford	3062	1566	1496	France	35	17	17	5	3	2	6	2	4	0	5	3	2	54	27	7	0	0	32	0.18	0.15	0.17	0.12	612	1.59	57	87	437	510	0.21	118	80	-1	5	
10	Aboubakar Kamara	25	794534400	Premier League	2018/2019	Forward	Fulham	687	468	219	France	13	4	1	3	1	2	0	0	1	1	2	1	1	16	9	7	2	0	0	39	0.39	0.39	0.19	0.82	229	2.1	43	53	344	0	26	37	412	-1	4
11	Adalberto Peñaranda	23	865033200	Premier League	2018/2019	Forward	Watford	0	0	0	Venezuela	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1							
12	Adam David Lallana	32	579222000	Premier League	2018/2019	Midfielder	Liverpool	465	189	276	England	13	2	3	0	0	0	0	0	6	1	5	2	1	1	1	0	0	0	0	39	233	36	465	0	19	379	344	-1	18						
13	Adam Masina	26	757468800	Premier League	2018/2019	Defender	Watford	1003	463	540	Italy	14	5	6	0	0	1	1	0	0	2	1	1	19	8	11	5	0	0	09	0.09	0.09	0	0	1	7	53	72	201	1003	0.45	397	155	143	20	
14	Adam Smith	29	672879600	Premier League	2018/2019	Defender	AFC Bournemouth	2073	1051	1022	England	25	12	13	1	1	1	0	1	0	9	5	4	30	10	20	6	1	0	09	0.04	0.04	0.09	0.20	73	1.3	69	83	296	2073	0.3	237	228	85	10	
15	Adama Diakhaby	24	836521200	Premier League	2018/2019	Forward	Huddersfield Town	551	345	206	France	12	4	2	0	0	0	0	0	1	1	0	16	7	9	1	0	0	0	0	2	61	34	46	551	0	16	332	359	-1	26					
16	Adama Traoré	24	822528000	Premier League	2018/2019	Midfielder	Wolverhampton Wanderers	890	315	575	Spain	29	2	6	1	0	1	1	1	0	0	6	4	2	11	2	9	1	0	0	2	0.1	0.1	0.1	0.16	890	1.11	81	31	890	890	0.1	169	152	-1	13
17	Ademola Lookman	22	877302000	Premier League	2018/2019	Forward	Everton	601	334	267	England	21	2	1	0	0	2	2	0	0	8	5	3	7	2	5	0	0	3	0	3	0	0	1	05	86	29	0	301	0	292	20	-1	17		
18	Adrian Mariappa	33	528678000	Premier League	2018/2019	Defender	Watford	1921	841	1080	Jamaica	26	8	12	0	0	0	0	0	0	6	3	3	29	10	19	3	0	0	0	1	36	66	74	640	0	14	396	414	94	21					
19	Adrián San Miguel del Castillo	33	536630400	Premier League	2018/2019	Goalkeeper	West Ham United	0	0	0	Spain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1							
20	Adrien Sebastian Perruchet Silva	31	605923200	Premier League	2018/2019	Midfielder	Leicester City	88	8	80	Portugal	2	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0	0	0	0	1	02	88	44	0	0	-1	-1	-1	15						

Figura 2 - Ficheiro CSV

Formatação do ficheiro

Aqui encontram-se alguns dos vários jogadores apresentados neste csv, com os seus vários atributos, sendo que alguns apresentam algumas irregularidades que irão necessitar de formatação.

Select & Alter Remove Meta-data				
Fields :				
#	Fieldname	Rename to	Length	Precision
1	full_name			
2	age			
3	birthday			
4	league			
5	season			
6	position			
7	Current_Club			
8	nationality			
9	goals_overall			

Figura 3 - Select Values

Aqui retirei vários dos atributos aos jogadores desnecessários, sendo que apenas considereei relevantes estes nove, considereei fazer um rename com a respetiva tradução mas visto que se trata de informação da liga inglesa, acabei por deixar as respetivos Fieldnames com os mesmos nomes.

Fields to compare on (no entries means: compare complete row)

#	Fieldname	Ignore case	
1	full_name	Y	
2	age	Y	
3	birthday	Y	
4	league	Y	
5	season	Y	
6	position	Y	
7	Current_Club	Y	
8	nationality	Y	
9	goals_overall	Y	

Figura 4 - Unique Rows

Fields

#	Field	Replace by value	Conversion mask (Date)	Set empty string?
1	full_name	ND		N
2	age	ND		N
3	birthday	ND		N
4	league	ND		N
5	season	ND		N
6	position	ND		N
7	Current_Club	ND		N
8	nationality	ND		N
9	goals_overall	ND		N

Figura 5 - Reple Nulls with ND

Posteriormente de modo a evitar repetições de linhas ou valores sem informação acabei por utilizar dois processos do kettle, ambos indicados respetivamente nas imagens acima, acabando por eliminar algumas linhas com informação desnecessária, para acabar de formatar a informação de forma pretendia.

Para acabar de formatar fiz operações entre strings, eliminando todos os números que poderiam estar nestas assim como espaços desnecessários, entre outros, para finalizar optei por fazer um sort rows nos nomes de modo a imprimir a informação destes por ordem alfabética.

String operations

Step name String operations

The fields to process:

#	In stream field	Out stream field	Trim type	Lower/Upper	Padding	Pad char	Pad Length	InitCap	Escape	Digits	Remove Special character
1	full_name		both	lower	none			S	None	remove	none
2	league		both	lower	none			S	None	remove	none
3	season		both	lower	none			S	None	none	none
4	position		both	lower	none			S	None	remove	none
5	Current_Club		both	lower	none			S	None	remove	none
6	nationality		both	lower	none			S	None	remove	none

Figura 6 - String Operations

Sort rows

Nome do Step

Sort rows

Sort directory

%%java.io.tmpdir%%

Navega...

TMP-file prefix

out

Sort size (rows in memory)

1000000

Free memory threshold (in %)

Compress TMP Files?

☒

Only pass unique rows? (verifies keys only)

☐

Fields:

#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?
1	full_name	S	N	N	0	N
2	age	N	N	N	0	N
3	birthday	N	N	N	0	N
4	league	N	N	N	0	N
5	season	N	N	N	0	N
6	position	N	N	N	0	N
7	Current_Club	N	N	N	0	N
8	nationality	N	N	N	0	N
9	goals_overall	N	N	N	0	N

Help

OK

Cancela

Obtem campos

Figura 7 - Sort Rows

Switch/ Case

Após ter feito a formatação dos dados optei por separar agora os jogadores conforme as suas posições, “Avançado”, “Defesa”, “Guarda-Redes” e “Médio”, para isso utilizei um Switch/Case seguido de um Filter Row com um ficheiro de texto como output, tendo por fim juntado todos os ficheiros de novo, o principal objetivo de ter realizado esta operação foi apenas para experimentar mais funcionalidades do kettle.

Switch / case

Step name: Switch / case

Field name to switch: position

Use string contains: ☐

Case value data type: String

Case value conversion mask:

Case value decimal symbol:

Case value grouping symbol:

#	Value	Target step
1	Forward	Filter rows
2	Defender	Filter rows 2
3	Midfielder	Filter rows 3
4	Goalkeeper	Filter rows 4

Default target step: Filter rows

Buttons: Help, OK, Cancela

Figura 8 - Switch/ case

Filter rows

Step name: Filter rows

Send 'true' data to step: Forward

Send 'false' data to step:

The condition:

☐ position = Forward (String)

Buttons: Help, OK, Cancela

Figura 9 - Filter Rows

Outputs

Posteriormente através do ficheiro de texto, optei por enviar o output do mesmo para uma base de dados e para um xml, como indicado na imagem seguinte.

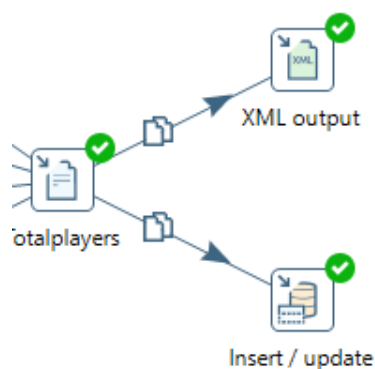


Figura 10 - Outputs

Carregamento para a base de dados

Inicialmente fiz a configuração de conexão à base de dados, na qual optei por utilizar o PostgreSQL por já ter trabalhado com ele.

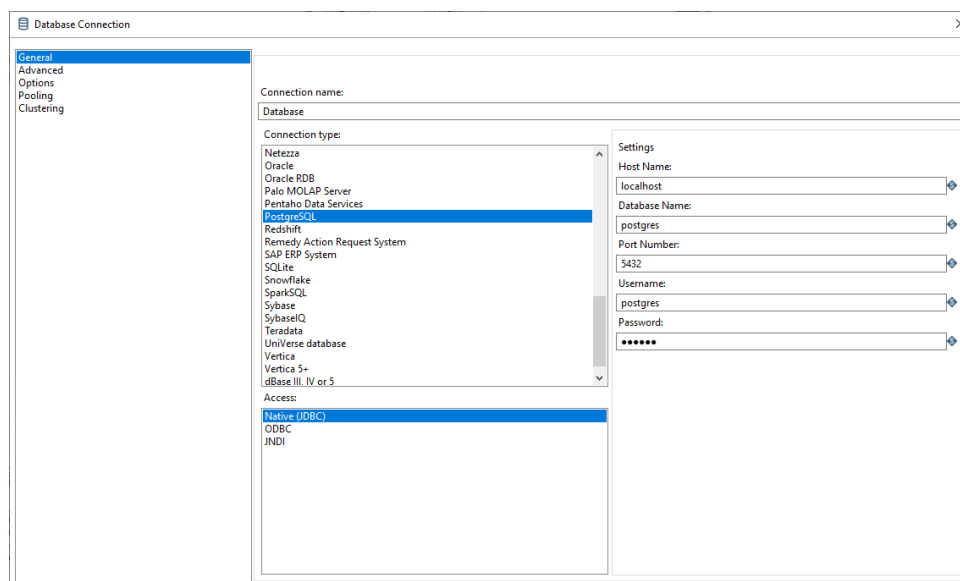


Figura 11 - Database Connection

Depois de ter configurado a base de dados procedi à configuração do Insert / update na qual igualei os atributos do ficheiro para a base de dados.

Step name: Insert / update

Connection: connection_postgres

Target schema: public

Target table: players

Commit size: 100

Don't perform any updates: ☐

The key(s) to look up the value(s):

#	Table field	Comparator	Stream field1	Stream field2
1	full_name	=	full_name	
2	age	=	age	
3	birthday	=	birthday	
4	league	=	league	
5	season	=	season	
6	position	=	position	
7	Current_Club	=	Current_Club	

Update fields:

#	Table field	Stream field	Update
1	full_name	full_name	Y
2	age	age	Y
3	birthday	birthday	Y
4	league	league	Y
5	season	season	Y
6	position	position	Y
7	Current_Club	Current_Club	Y
8	nationality	nationality	Y
9	goals_overall	goals_overall	Y

Figura 12 - Insert / update

Ficheiro XML

Nome do Step: XML output

Filename: \${Internal.Entry.Current.Directory}/Output.xml

Do not create file at start: ☐

Pass output to servlet: ☐

Extension: xml

Include stepnr in filename?: ☐

Include date in filename?: ☐

Include time in filename?: ☐

Specify Date time format: ☐

Date time format:

Show filename(s):...

Add filenames to result: ☐

Figura 13 - XML

Kettle job

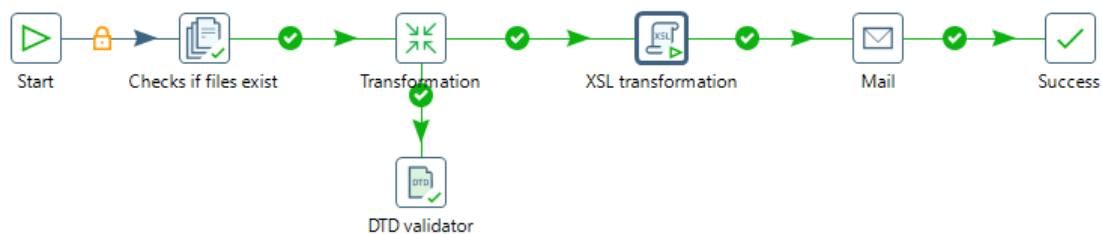


Figura 14 – Job

Inicialmente verificámos se o csv existe, caso exista então ocorre a transformação mencionada anteriormente, tendo o xml de ser validado por um ficheiro DTD.

Ficheiro DTD

```

1 <!ELEMENT Rows (Row)+>
2 <!ELEMENT Row (full_name,age,birthday,league,season,position,Current_Club,nationality,goals_overall)>
3 <!ELEMENT full_name (#PCDATA)>
4 <!ELEMENT age (#PCDATA)>
5 <!ELEMENT birthday (#PCDATA)>
6 <!ELEMENT league (#PCDATA)>
7 <!ELEMENT season (#PCDATA)>
8 <!ELEMENT position (#PCDATA)>
9 <!ELEMENT Current_Club (#PCDATA)>
10 <!ELEMENT nationality (#PCDATA)>
11 <!ELEMENT goals_overall (#PCDATA)>

```

Figura 15 - DTD file

Output de XSL para HTML

Criamos um stylesheet XSL simples que vai permitir que seja criada uma tabela HTML para apresentar os dados.

```
<?xml version="1.0" encoding="UTF-8"?>
<!--Diogo Rocha 16 LESI ISI-->
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
  - <xsl:template match="/">
    - <html>
      - <body>
        - <table display-align="center" cellpadding="1" cellspacing="1" width="100%" table-layout="fixed" bgcolor="LightSlateGray" border="1">
          - <tr>
            <td style="font-weight:bold; align="center" colspan="10">Jogadores da premier league</td>
          </tr>
          - <tr bgcolor="Moccasin">
            <th> full_name </th>
            <th> age </th>
            <th> birthday </th>
            <th> league </th>
            <th> season </th>
            <th> position </th>
            <th> Current_Club </th>
            <th> nationality </th>
            <th> goals_overall </th>
          </tr>
          - <xsl:for-each select="Rows/Row">
            - <tr>
              - <xsl:attribute name="bgcolor">
                - <xsl:choose>
                  <!-- Varia cores que dependeram da posição do jogador -->
                  <xsl:when test="position[.='Forward']">LightGrey</xsl:when>
                  <xsl:when test="position[.='Defender']">#FFD700</xsl:when>
                  <xsl:when test="position[.='Goalkeeper']">#FFFACD</xsl:when>
                  <xsl:when test="position[.='Midfielder']">#228B22</xsl:when>
                  <xsl:otherwise>PowderBlue</xsl:otherwise>
                </xsl:choose>
              </xsl:attribute>
              - <td>
                <xsl:value-of select="full_name"/>
              </td>
              - <td>
                <xsl:value-of select="age"/>
              </td>
              - <td>
                <xsl:value-of select="birthday"/>
              </td>
              - <td>
                <xsl:value-of select="league"/>
              </td>
              - <td>
                <xsl:value-of select="season"/>
              </td>
              - <td>
                <xsl:value-of select="position"/>
              </td>
              - <td>
                <xsl:value-of select="Current_Club"/>
              </td>
              - <td>
                <xsl:value-of select="nationality"/>
              </td>
              - <td>
                <xsl:value-of select="goals_overall"/>
              </td>
            </tr>
          </xsl:for-each>
        </table>
      </body>
    </html>
  </xsl:template>
</xsl:stylesheet>
```

Figura 16- XSL

XSL Transformation

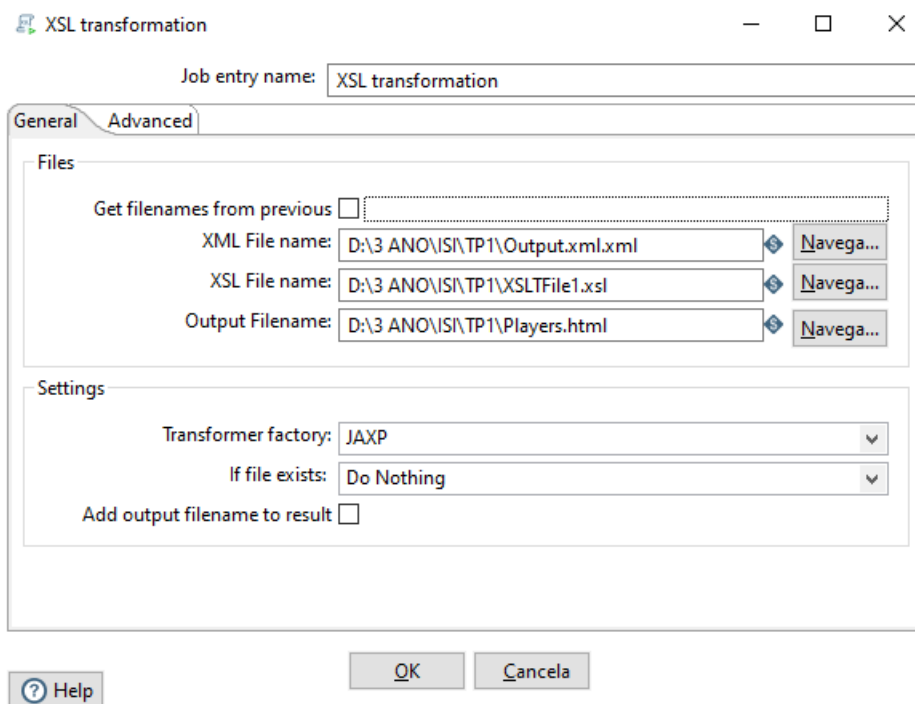


Figura 17- XSL transformation

HTML

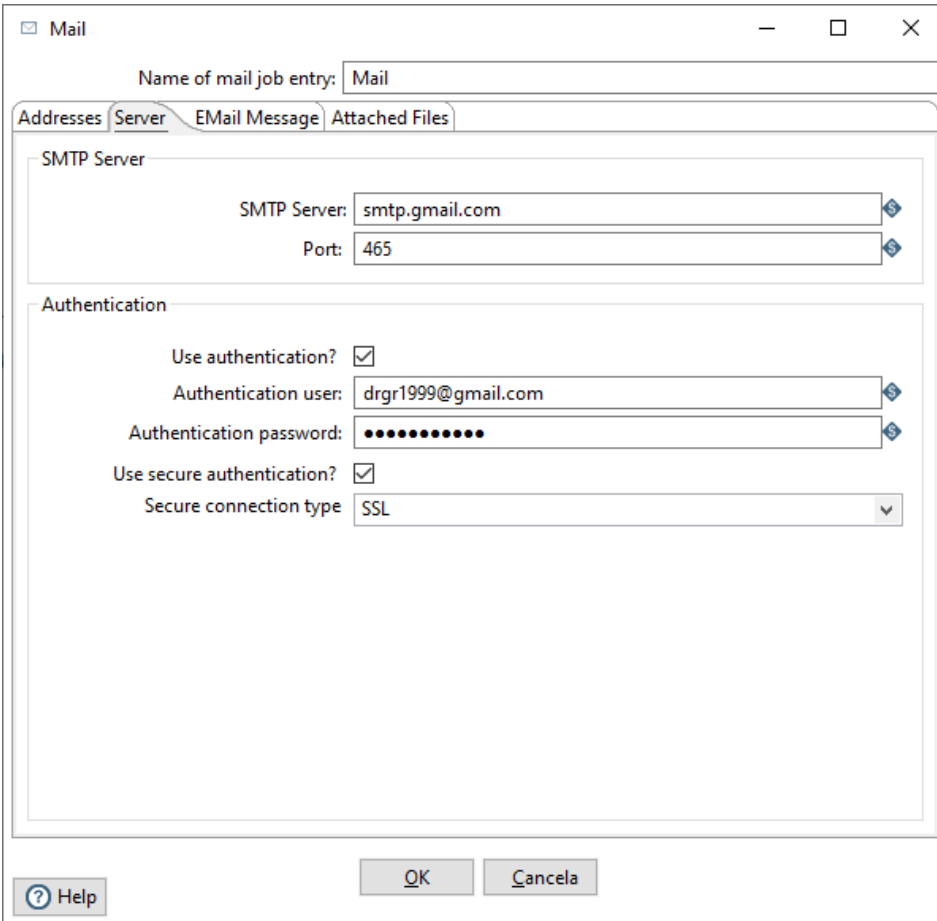
Output de uma pequena parte de HTML

Jogadores da premier league									
full_name	age	birthday	league	season	position	Current_Club	nationality	goals_overall	
Aaron Lennon	33	545526000	Premier League	2018/2019	Midfielder	Burnley	England	1	
Aaron Mooy	30	653353200	Premier League	2018/2019	Midfielder	Huddersfield Town	Australia	3	
Aaron Ramsey	29	662169600	Premier League	2018/2019	Midfielder	Arsenal	Wales	4	
Aaron Wan-Bissaka	22	880502400	Premier League	2018/2019	Midfielder	Crystal Palace	England	0	
Abdelhamid Sabiri	23	849139200	Premier League	2018/2019	Midfielder	Huddersfield Town	Morocco	0	
Abdoulaye Doucoura	27	725846400	Premier League	2018/2019	Midfielder	Watford	France	5	
Adam David Lallana	32	579222000	Premier League	2018/2019	Midfielder	Liverpool	England	0	
Adama Traoré	24	822528000	Premier League	2018/2019	Midfielder	Wolverhampton Wanderers	Spain	1	
Adrien Sebastian Perruchet Silva	31	605923200	Premier League	2018/2019	Midfielder	Leicester City	Portugal	0	
Ainsley Maitland-Niles	23	872809200	Premier League	2018/2019	Midfielder	Arsenal	England	1	
Alex Oxlade-Chamberlain	27	745369200	Premier League	2018/2019	Midfielder	Liverpool	England	0	

Figura 18- HTML

Email

Configuração do email.



The screenshot shows a 'Mail' configuration window with the following fields and options:

- Name of mail job entry:** Mail
- Addresses** | **Server** | **Email Message** | **Attached Files**
- SMTP Server**
 - SMTP Server: smtp.gmail.com
 - Port: 465
- Authentication**
 - Use authentication? ☒
 - Authentication user: drgr1999@gmail.com
 - Authentication password: [masked]
 - Use secure authentication? ☒
 - Secure connection type: SSL
- Buttons:** ? Help, OK, Cancela

Figura 19- Mail

Inicialmente tive algumas dificuldades a conseguir enviar emails, devido ao facto de ser necessário desativar algumas proteções que o email possui, mas após desativar já recebo os emails com o report do trabalho.

Output Email

teste

Job:

JobName : KETTLE

Directory : /

JobEntry : Mail

Message date: 2020/11/13 19:07:18.522

Previous results:

Job entry Nr : 2

Errors : 0

Lines read : 0

Lines written : 1

Lines input : 0

Lines output : 0

Lines updated : 0

Lines rejected : 0

Script exist status : 0

Result : true

Path to this job entry:

KETTLE

KETTLE :: start : Start of job execution (2020/11/13 19:07:18.247)

KETTLE :: Start : start : Start of job execution (2020/11/13 19:07:18.248)

KETTLE :: Start : [nr=0, errors=0, exit_status=0, result=true] : Job execution finished (2020/11/13 19:07:18.248)

KETTLE :: Checks if files exist : Followed unconditional link : Start of job execution (2020/11/13 19:07:18.249)

KETTLE :: Checks if files exist : [nr=0, errors=0, exit_status=0, result=true] : Job execution finished (2020/11/13 19:07:18.249)

KETTLE :: Transformation : Followed link after success : Start of job execution (2020/11/13 19:07:18.249)

KETTLE :: Transformation : [nr=2, errors=0, exit_status=0, result=true] : Job execution finished (2020/11/13 19:07:18.518)

KETTLE :: XSL transformation : Followed link after success : Start of job execution (2020/11/13 19:07:18.518)

KETTLE :: XSL transformation : [nr=2, errors=0, exit_status=0, result=true] : Job execution finished (2020/11/13 19:07:18.519)

KETTLE :: Mail : Followed link after success : Start of job execution (2020/11/13 19:07:18.519)

Figura 20 - Output Email

Conclusão

De uma forma geral, este trabalho ajudou-me a desenvolver as minhas capacidade de trabalho individual, no desenvolvimento de uma transformação com o uso de uma ferramenta ETL que neste caso foi Open Source, Kettle. Ajudou-me também por sua vez a perceber melhor como funciona melhor esta ferramenta e o quão útil e completa é.

Bibliografia

- Aulas lecionadas em ISI – Integração de Sistemas de Informação;
- Documentos fornecidos pelo professor no moodle, trabalhos previamente realizados, assim como livros em formato pdf's e vídeos;