

Integração de Sistemas de Informação

1º Trabalho prático

Gonçalo Costa, N°26024

Docente: Óscar Ribeiro

Índice

Introdução.....	4
Problema.....	5
Estratégia Utilizada	6
Softwares Utilizados	6
Transformações.....	7
Transformações Desenvolvidas	7
Jobs	16
Jobs Desenvolvidos.....	16
Email.....	18
Node-RED	21
Vídeo demonstrativo	23
Conclusão.....	24
Bibliografia/Web grafia	25

Índice de Figuras

Figura 1: Transformação htmlAsAmbient	7
Figura 2: Transformação 1	8
Figura 2.1: Passos de enviar dados para a base de dados	8
Figura 2.2: Conexão com a Base de Dados SQL.....	9
Figura 2.3: Filter Rows onde expressão regular	9
Figura 2.4: If field value is null Transformação 1	10
Figura 2.5: If field value is null	10
Figura 2.6: Verificação se a música tem link_url.....	10
Figura 2.7: Verificação se a música tem mais de 1 artista.....	10
Figura 2.8: Renomeação dos campos com o Select Values	11
Figura 3: Renomeação dos campos com o Select Values	12
Figura 3.1: Acesso a Base de dados pretendida e query SQL	12
Figura 3.2: Ordenação pelos BPM de forma decrescente	13
Figura 3.3: Script com o código	13
Figura 3.4: Regex Evaluation	14
Figura 3.5: Verificação se passou ou não	14
Figura 3.6: Verificação de Cantor	15
Figura 4: Job 1	16
Figura 4.1: Eliminação do ficheiro XML.....	16
Figura 4.2: Evocação da Transformação	16
Figura 4.3: Verificações Necessárias	17
Figura 4.4: XSL Transformation	17
Figura 4.5: Transformation HTML_ambient	17
Figura 4.6: Envio de Email.....	17
Figura 4.7: Log de Sucesso!	18
Figura 5: Email que chegou à caixa de entrada do destinatário	18
Figura 5.1: Configuração de Email (destination e sender address).....	19
Figura 5.2: Configuração de Email (server)	19
Figura 5.3: Configuração de Email (mensagem de email)	20
Figura 6: Esquema realizado no Node-RED.....	21
Figura 6.1: Chamada da function e conexão com SQL	21
Figura 6.2: Query SQL.....	22
Figura 6.4: Resultado do Esquema de Node-RED.....	22
Figura 8: QR Code	23

Introdução

Com este trabalho para a unidade curricular de Integração de Sistemas de Informação (ISI) pretende-se focar a aplicação 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. Ferramentas complementares como node-RED, poderão também ser exploradas e integradas nos processos.

Problema

A análise de dados de streaming de música, especialmente de plataformas como o Spotify, fornece insights valiosos sobre tendências de consumo, preferências dos utilizadores e desempenho de artistas. Contudo, a natureza volumosa e variada dos dados contidos em ficheiros CSV representa um desafio significativo para os analistas. Estes ficheiros não apenas incluem informações básicas, como nomes de artistas e faixas, mas também métricas complexas, como o número de streams, classificações e outros indicadores de desempenho.

Para abordar estas questões, um processo de ETL (Extract, Transform, Load) eficaz torna-se indispensável. Este processo não só ajudará a otimizar o fluxo de dados, mas também garantirá a qualidade e integridade da informação analisada. A implementação de ferramentas de ETL apropriadas permitirá a automação de tarefas repetitivas, a redução de erros manuais e, em última instância, a capacidade de gerar relatórios e análises em tempo real.

Assim, este projeto visa desenvolver uma solução robusta que possa enfrentar estes desafios e facilitar a exploração e interpretação dos dados de streaming de música.

Estratégia Utilizada

Neste projeto, a estratégia de ETL (Extract, Transform, Load) foi implementada utilizando o Pentaho Data Integration (PDI), uma ferramenta poderosa e flexível para manipulação de dados.

A entrada de dados ocorre a partir de um ficheiro CSV, que contém todas as informações necessárias para a transformação. Durante esta fase inicial, são realizadas verificações para identificar se existem valores nulos. Nos casos em que isso acontece, são atribuídos valores padrão para assegurar a continuidade do processamento.

É igualmente importante verificar a legibilidade dos caracteres apresentados. Isso implica a deteção de caracteres especiais que possam causar erros na leitura do conteúdo dos ficheiros. Para contornar essas situações, outros caracteres ou expressões são assumidos, garantindo a integridade dos dados.

Ao longo de todas as transformações, diversas operações de filtragem e ordenação são aplicadas. Através destas operações, garantimos que apenas os valores relevantes sejam utilizados e que não existam anomalias nos ficheiros, como dados duplicados ou inconsistentes.

Finalmente, os dados tratados são armazenados num servidor SQL, permitindo a sua consulta e leitura para futuras transformações necessárias. Esta abordagem não só facilita a gestão dos dados, mas também assegura que as informações estejam organizadas e prontas para análises subsequentes.

Softwares Utilizados

- Pentaho Kettle
- SQL Server
- Node-Red

Transformações

No Pentaho Kettle, as transformações são usadas para realizar processos de ETL, ou seja, extração, transformação e carga de dados. Na fase de extração, os dados podem ser obtidos de várias fontes, como ficheiros CSV, Excel, bases de dados, utilizando passos como o CSV File Input ou o Table Input.

Transformações Desenvolvidas

Transformação *htmlAsAmbient*



Figura 1: Transformação *htmlAsAmbient*

A figura apresenta a transformação denominada *htmlAsAmbient*, composta por três etapas principais.

A primeira é o *Text file input*, onde um ficheiro de texto é lido como fonte de dados. Após isso, a operação *Group by* agrupa os dados com base em uma ou mais colunas específicas, permitindo a agregação de informações. Por fim, a etapa *SetStringAmbient* parece aplicar uma transformação nos dados agrupados, provavelmente definindo ou ajustando o valor de uma string relacionada ao ambiente HTML.

Essa transformação permite organizar e manipular os dados provenientes de um ficheiro de texto, ajustando-os de forma adequada para serem utilizados no contexto de HTML, como parte do processo de preparação do conteúdo para ser enviado ou exibido.

Transformação 1

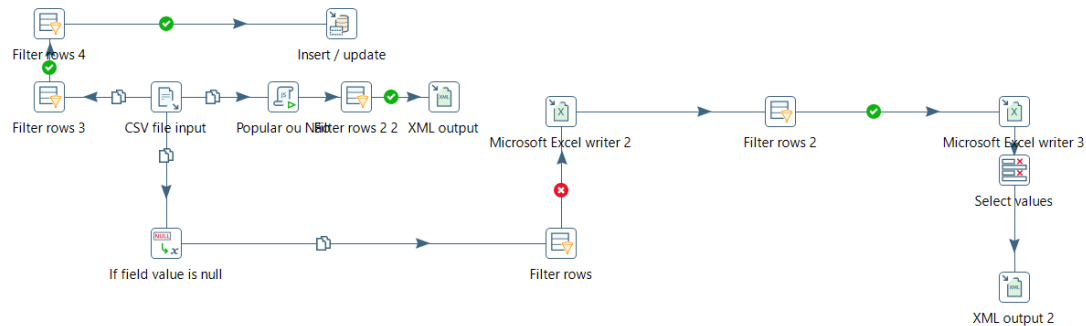


Figura 2: Transformação 1

Na transformação acima apresentada, existe a leitura de informação a partir de um ficheiro CSV (ficheiro inicial que contém a informação inicial).

Nesta transformação, são efetuadas diferentes operações. Na primeira, foram realizados dois Filter Rows, para verificar a existência de caracteres especiais em 2 campos diferentes (track_name e artist_name), após essa avaliação, são inseridos na base de dados SQL (Insert/Update).

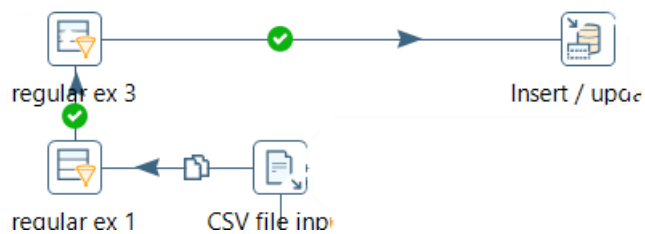


Figura 2.1: Passos de enviar dados para a base de dados

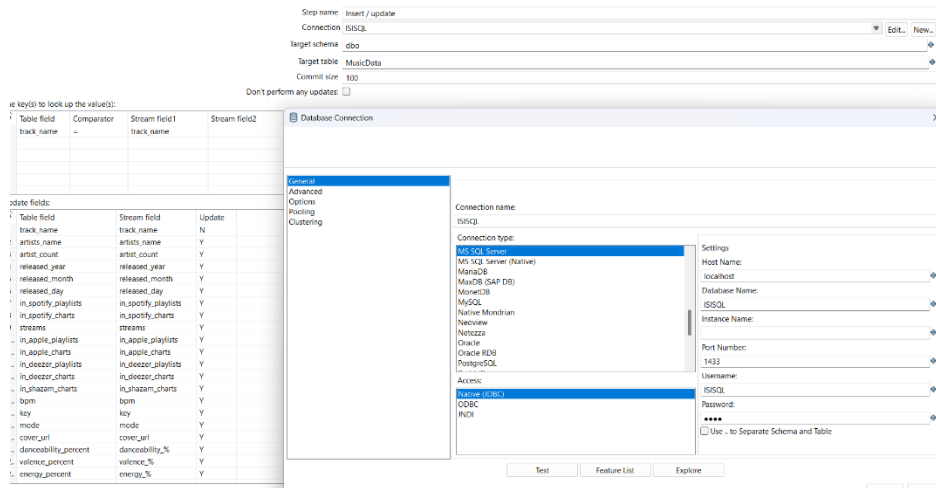


Figura 2.2: Conexão com a Base de Dados SQL

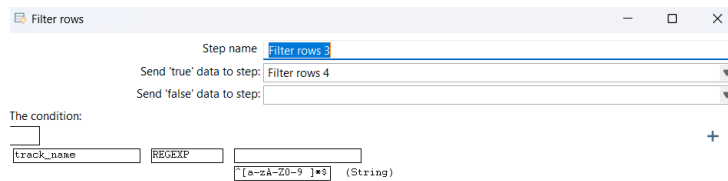


Figura 2.3: Filter Rows onde expressão regular

Aqui, verifica-se se o no campo pretendido, apenas existem caracteres que podem ser lidos, ou seja, sem caracteres especiais.

Além disso, são feitas mais alterações/configurações dos dados para conseguir feita a sua manipulação de forma correta. Um desses exemplos é a utilização do passo “If field value is null”, onde são atribuídos valores assumir se o valor do campo for nulo.

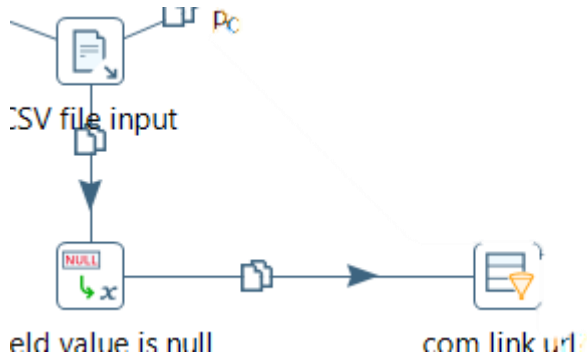


Figura 2.4: If field value is null Transformação 1

#	Field	Replace by value	Conversion mask (Date)	Set empty string?
1	track_name			N
2	artists_name			N
3	artist_count			N
4	released_year			N
5	released_month			N
6	released_day			N
7	in_spotify_playlists			N
8	in_spotify_charts			N
9	streams			N
1.	in_apple_playlists			N
1.	in_apple_charts			N
1.	in_deezer_playlists			N
1.	in_deezer_charts			N
1.	in_shazam_charts	-1		N
1.	bpm			N
1.	key	NA		N
1.	mode			N

Figura 2.5: If field value is null

Ao vermos esta imagem, podemos ver que no campo *in_shazam_charts*, vai ser atribuído o valor -1 e no campo *KEY* o valor NA.

Ao longo desta transformação são realizadas mais algumas filtrações, como por exemplo, se contém “link_url” ou se a música tem mais de 1 artista(contém “,” no campo).

cover_url = (String)

Figura 2.6: Verificação se a música tem link_url

artists_name CONTAINS (String)

Figura 2.7: Verificação se a música tem mais de 1 artista

Alguns caracteres podem não ser lidos com sucesso, como por exemplo o “%”, com um *Select Value* é possível renomear o campo e assim resolver esse problema.



Figura 2.8: Renomeação dos campos com o *Select Values*

Após isso, se tudo gravado com sucesso, o conteúdo desta transformação é armazenado num ficheiro do tipo XML.

Transformação 2

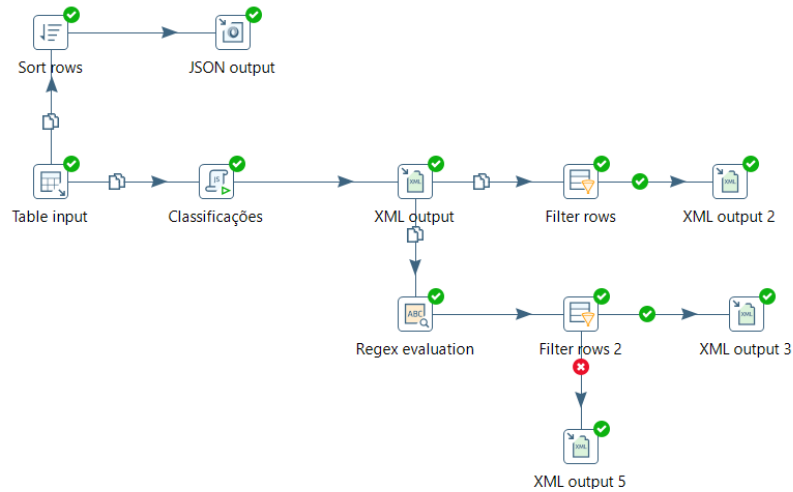


Figura 3: Renomeação dos campos com o Select Values

Esta transformação começa com a leitura da base de dados que foi gravada na Transformação 1.

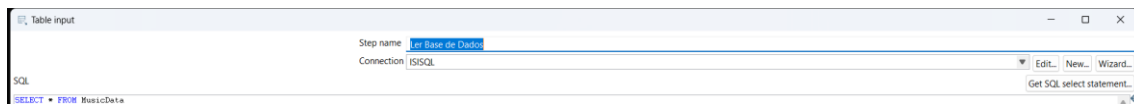


Figura 3.1: Acesso a Base de dados pretendida e query SQL

A partir desses dados lidos, são executadas algumas operações para trabalhar com os dados

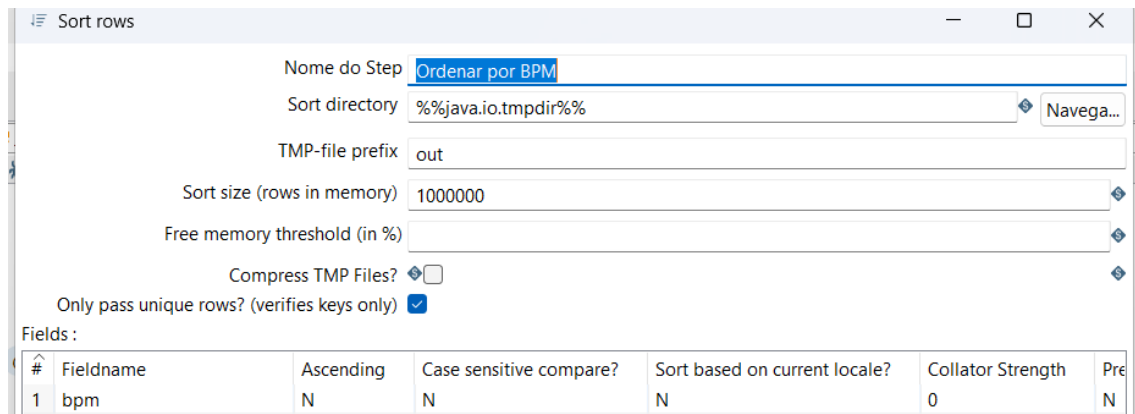


Figura 3.2: Ordenação pelos BPM de forma decrescente

Na imagem acima, recorrendo a um Sort Rows, tendo em conta o seu valor de BPM, as músicas são ordenadas de forma decrescente e armazenadas num ficheiro JSON.

Também foram realizadas operações recorrendo a código JavaScript em que são analisadas e atribuídas novas colunas com as conclusões retiradas a partir do código,

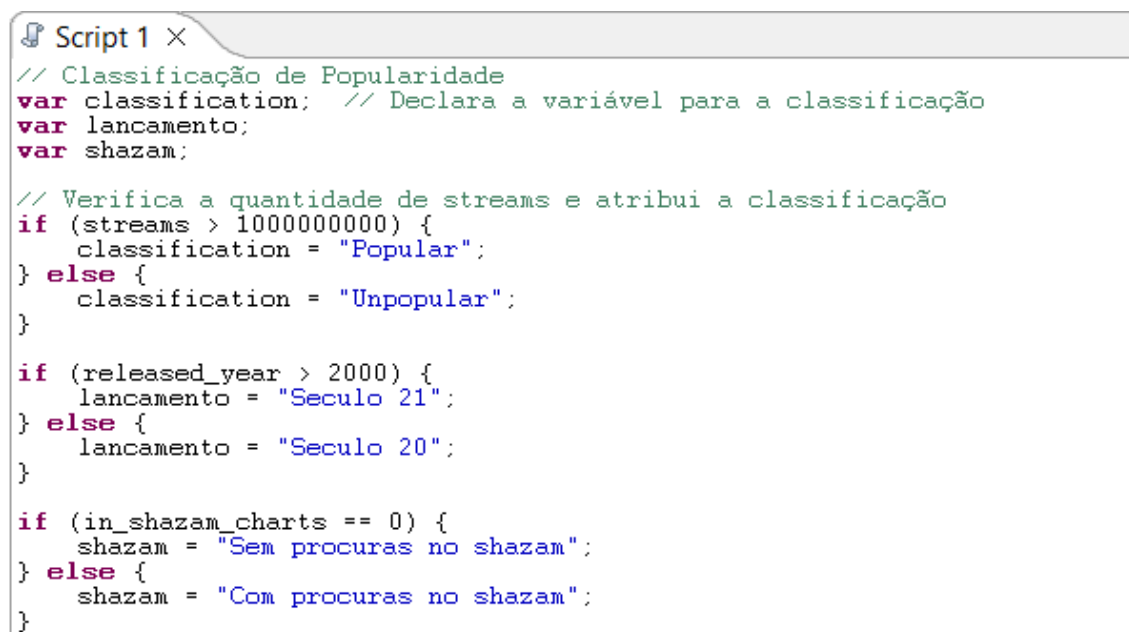


Figura 3.3: Script com o código

A partir deste código, são criadas, para cada música, 3 novas variáveis, classification, lançamento e shazam, tendo em conta as conclusões retiradas a partir do algoritmo e são guardados num ficheiro XML.

Após isso, recorrendo a um Regex Evaluation, verifica-se, se o campo Artists_name não tem espaços

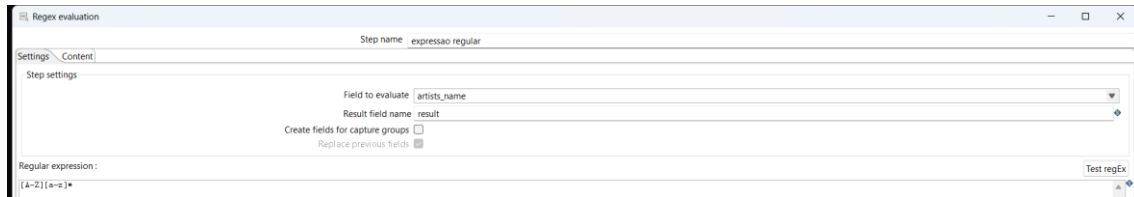


Figura 3.4: Regex Evaluation

O resultado deste Regex Evaluation é uma avaliação, se passou atribui um “Y”, se não passou, atribui um “N” e com isso serão filtrados os que passaram ou não.

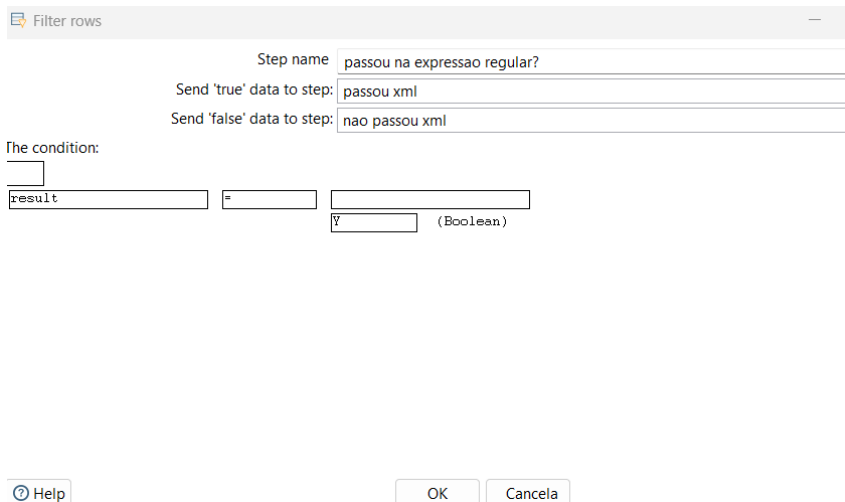
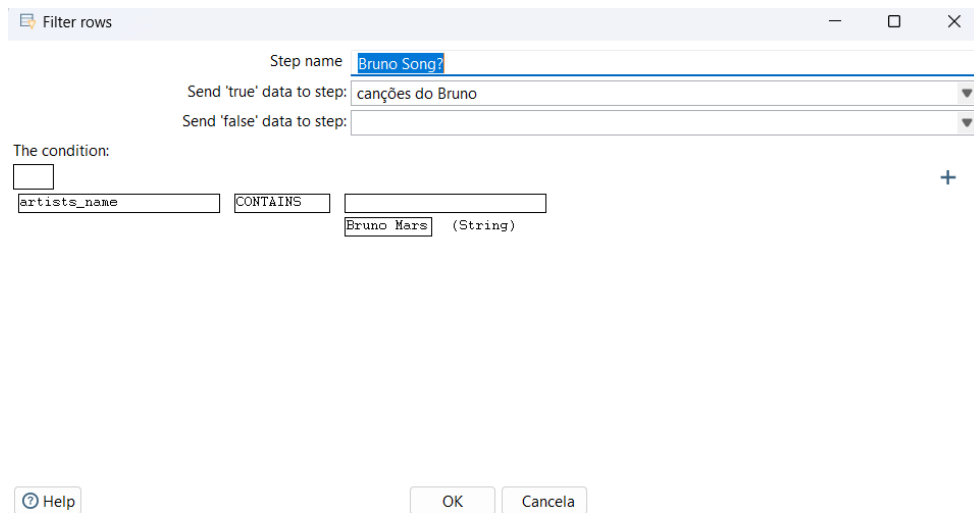


Figura 3.5: Verificação se passou ou não

Aqui analisa-se o resultado da avaliação, separa-se tendo em conta o resultado da avaliação da expressão regular.

Se passou vai para um ficheiro XML e se não passou para outro



Filter rows

Step name: Bruno Song?

Send 'true' data to step: canções do Bruno

Send 'false' data to step:

The condition:

☐ ☐

artists_name CONTAINS Bruno Mars (String)

Help OK Cancela

Figura 3.6: Verificação de Cantor

Aqui verifica-se se no campo Artists_name contém “Bruno Mars” e caso sim, armazena-se num ficheiro de XML.

Jobs

Os jobs permitem incluir várias tarefas, como executar transformações, transferir ficheiros, enviar emails ou executar scripts. Um exemplo de tarefa comum é o passo Start, que define o ponto de início de um job. A partir daí, podem-se configurar passos como o Transformation, que executa uma transformação previamente definida. Um job pode ter condicionais de fluxo, como o Success ou Failure, que permitem tomar decisões com base no resultado da execução de uma tarefa anterior.

Jobs Desenvolvidos

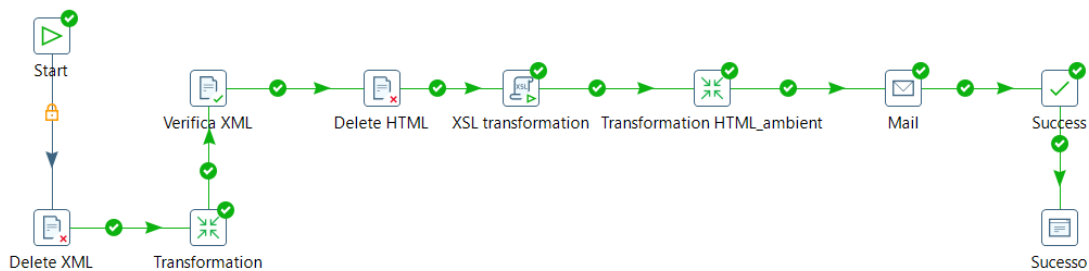


Figura 4: Job 1

Este JOB inicia com um processo de preparação para o tratamento de ficheiros XML e HTML. O primeiro passo é verificar que o ficheiro XML está eliminado.



Figura 4.1: Eliminação do ficheiro XML

Após essa verificação, é realizada uma transformação (a Transformação 1) onde é gerado um Ficheiro XML que vai ser utilizado futuramente.



Figura 4.2: Evocação da Transformação

Após isso, o mesmo ficheiro XML é verificado para garantir que está no formato correto e que a sua estrutura é válida. Este processo é fundamental para assegurar que as operações subsequentes possam ser realizadas sem problemas. Após a verificação do XML, ocorre uma igual para o ficheiro HTML, eliminando ficheiros antigos que poderiam interferir no novo processamento.



Figura 4.3: Verificações Necessárias

Posteriormente, ocorre a transformação XSL, que converte o ficheiro XML num ficheiro HTML formatado de acordo com uma folha de estilo predefinida. Esta transformação é fundamental para a apresentação correta dos dados extraídos do XML.

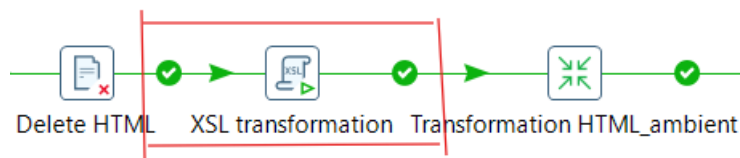


Figura 4.4: XSL Transformation

Após a transformação XSL, é realizada uma nova transformação no ficheiro HTML, denominada "Transformation HTML_ambient". Esta fase parece ajustar ou refinar o ambiente do ficheiro HTML, possivelmente aplicando formatações adicionais ou realizando ajustes nos dados antes do ficheiro ser enviado.

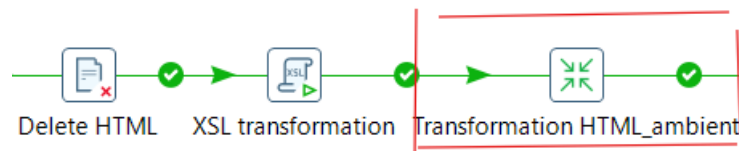


Figura 4.5: Transformation HTML_ambient

Com o HTML preparado, o próximo passo é o envio de um e-mail, que pode conter o ficheiro HTML gerado ou simplesmente uma notificação sobre o sucesso ou falha do processo. Após o envio do e-mail, o JOB é concluído com sucesso, registrando que todo o processo foi finalizado corretamente.



Figura 4.6: Envio de Email

Após todo o processo, é impresso na tela um log de sucesso.

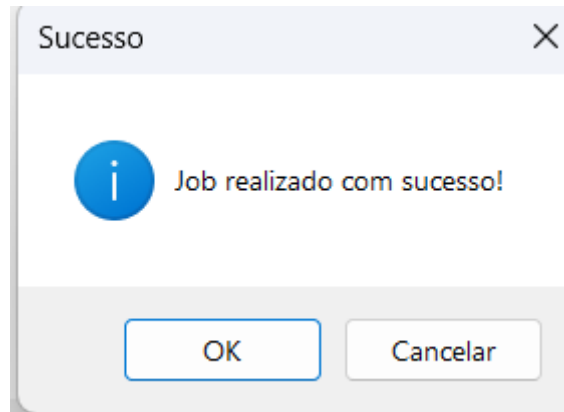


Figura 4.7: Log de Sucesso!

Email


 Gonçalo Cardoso Ferreira da Costa
 Para:  Gonçalo Cardoso Ferreira da Costa

Spotify Most Streamed Songs

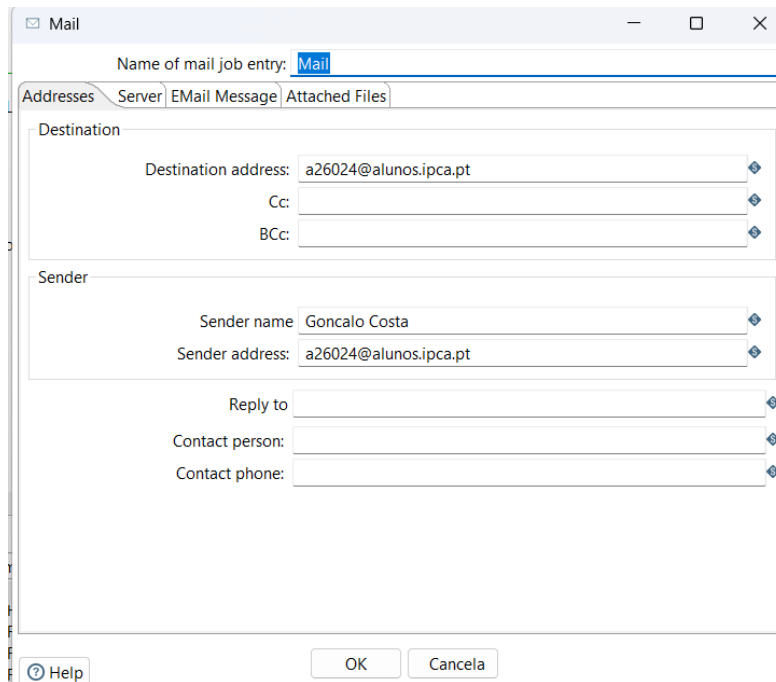
Riton, Nightcrawlers, Mufasa & Hypeman, Dopamine

Track Name	Artists Name	Artist Count	Released Year	Released Month	Released Day	In Spotify Playlists	In Spotify Charts
Sprinter	Dave, Central Cee	2	2023	6	1	2186	91
Ella Baila Sola	Eslabon Armado, Peso Pluma	2	2023	3	16	3090	50
Classy 101	Feid, Young Miko	2	2023	3	31	2610	40
Peso Pluma: Bzrp Music Sessions, Vol. 55	Bizarrap, Peso Pluma	2	2023	5	31	1313	40
Popular (with Playboi Carti & Madonna) - The Idol Vol.	The Weeknd, Madonna, Playboi Carti	3	2023	6	2	1945	87

Figura 5: Email que chegou à caixa de entrada do destinatário

Como pode ser visto pela imagem acima, o email chegou ao endereço de destino com sucesso.

Isto deve-se à sua correta configuração



Mail

Name of mail job entry: Mail

Addresses Server EMail Message Attached Files

Destination

Destination address: a26024@alunos.ipca.pt

Cc:

Bcc:

Sender

Sender name: Goncalo Costa

Sender address: a26024@alunos.ipca.pt

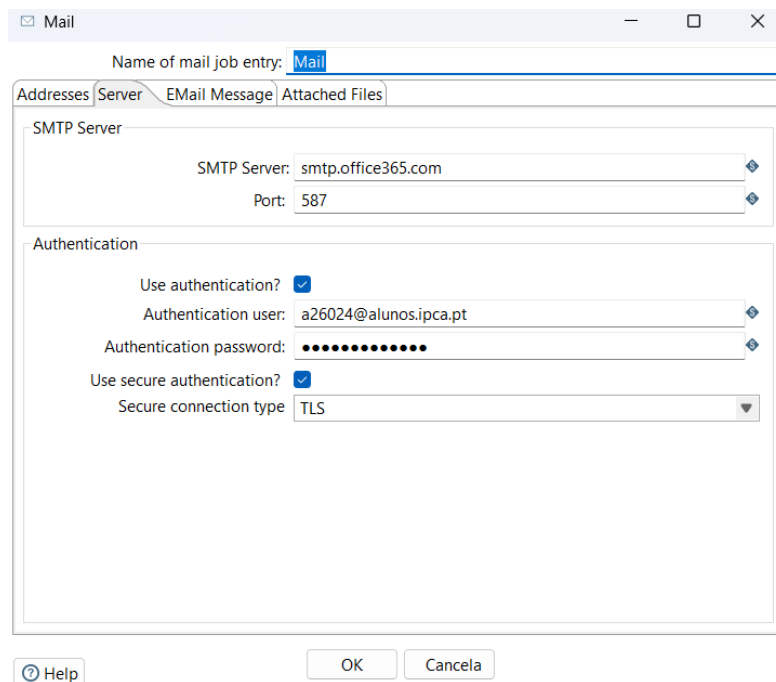
Reply to:

Contact person:

Contact phone:

Help OK Cancela

Figura 5.1: Configuração de Email (destination e sender address)



Mail

Name of mail job entry: Mail

Addresses Server EMail Message Attached Files

SMTP Server

SMTP Server: smtp.office365.com

Port: 587

Authentication

Use authentication? ☒

Authentication user: a26024@alunos.ipca.pt

Authentication password:

Use secure authentication? ☒

Secure connection type: TLS

Help OK Cancela

Figura 5.2: Configuração de Email (server)

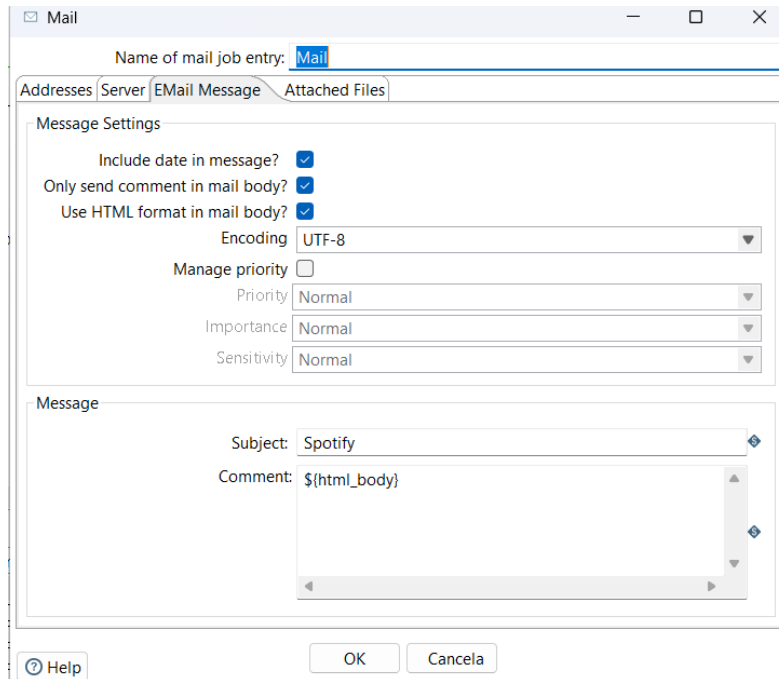


Figura 5.3: Configuração de Email (mensagem de email)

Node-RED

O Node-RED é uma ferramenta de desenvolvimento baseada em fluxos que permite conectar dispositivos, APIs e serviços online de maneira fácil e visual. Ele foi originalmente desenvolvido pela IBM e é amplamente utilizado para Internet das Coisas (IoT), automação e integração de sistemas.

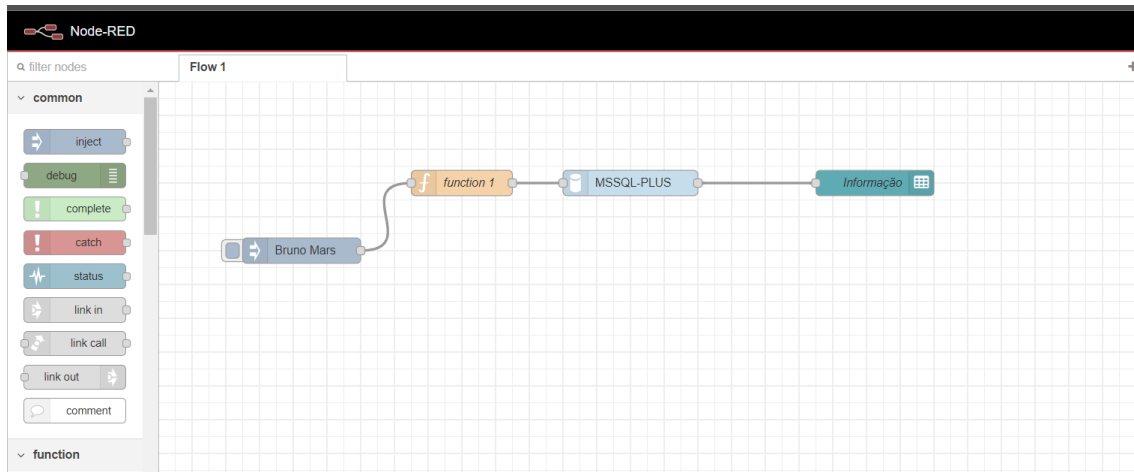


Figura 6: Esquema realizado no Node-RED

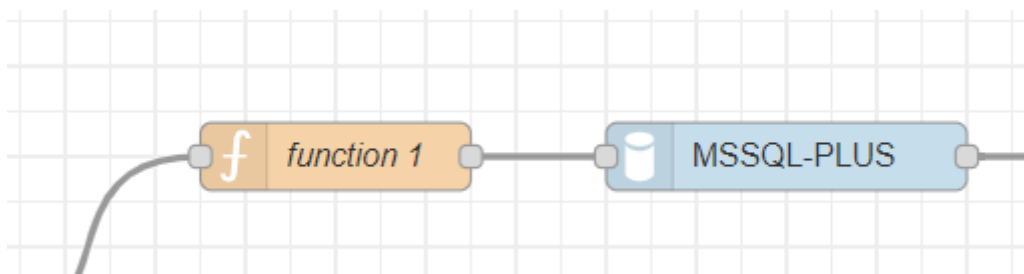


Figura 6.1: Chamada da function e conexão com SQL

Aqui, é chamada uma function com uma query para procurar na Base de Dados SQL informações pretendidas pelo utilizador.

A conexão com o SQL é feita pelo MSSQL-PLUS apresentado na imagem

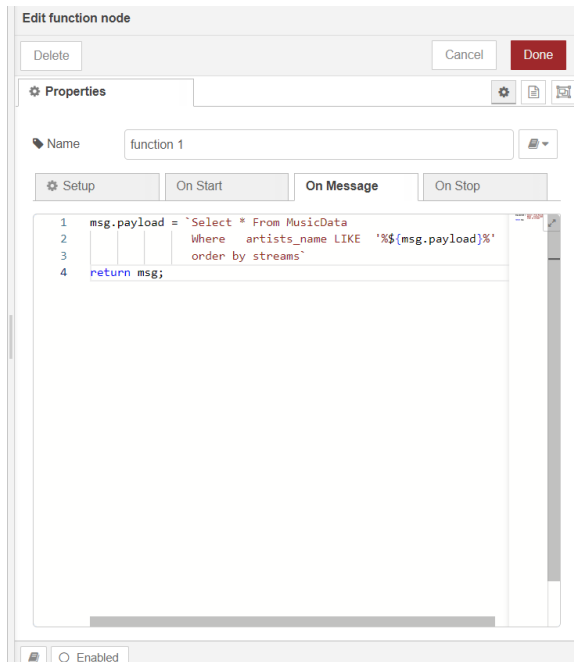


Figura 6.2: Query SQL

Musicas de Cantor																								
Tabela de informação																								
track_name	artists_name	artist_id	r_year	r_month	r_week	i_year	i_month	i_week	s_year	s_month	s_week	b_year	b_month	b_week	key	m_year	d_year	v_year	e_year	a_year	i_year			
Talking To The Moon	Bruno Mars	1	2010	10	4	true	true	10629...	5	0	862	0	0	146	C#	Minor	52	7	61	51	0			
Locked Out Of Heav...	Bruno Mars	1	2012	12	5	true	true	14813...	0	0	356	0	0	144	F	Major	73	87	70	6	0			
Just The Way You Are	Bruno Mars	1	2010	1	1	true	true	16414...	82	0	2946	0	0	109	F	Major	63	46	85	1	0			
When I Was Your Man	Bruno Mars	1	2012	12	5	true	true	16611...	0	0	806	0	0	145		Major	60	43	27	94	0			

Figura 6.4: Resultado do Esquema de Node-RED

Aqui, está apresentado o resultado de todo o processo apresentado anteriormente

Vídeo demonstrativo

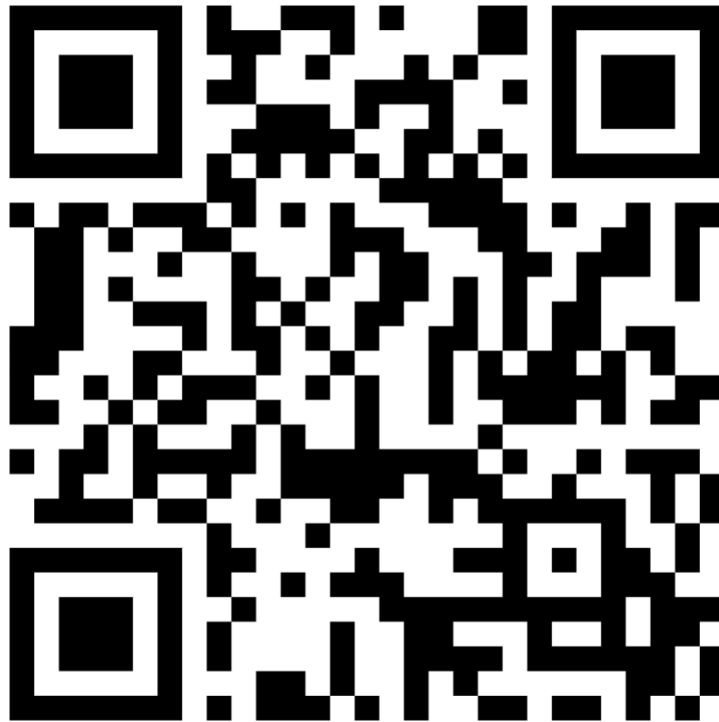


Figura 8: QR Code

Ao entrar neste QR Code, será redirecionado para um perfil, escolha a opção vídeo para o ver.

Conclusão

Neste contexto, conclui-se que o trabalho permitiu consolidar os conhecimentos relacionados com os processos de ETL, desde a extração, transformação e carga de dados, até à aplicação prática de ferramentas como o Pentaho Kettle. A integração de ferramentas complementares, como o Node-RED e o Home Assistant, foi uma oportunidade para explorar novas tecnologias e potenciar a automação de processos.

Os desafios enfrentados, como o tratamento de grandes volumes de dados e a manipulação de diferentes formatos (XML, JSON, bases de dados), foram superados através da utilização de técnicas adequadas, como o uso de expressões regulares, normalização de dados. A utilização de processos de logging e visualização dos resultados também se revelou uma mais-valia para monitorizar e validar a eficiência dos processos desenvolvidos.

Em suma, este trabalho proporcionou uma experiência prática significativa, permitindo a exploração de tecnologias atuais no contexto da Integração de Sistemas de Informação, reforçando a capacidade de desenvolver soluções escaláveis e eficientes.

Bibliografia/Web grafia

- Material disponibilizado pelo professor da unidade curricular
- Sebenta Pentaho Kettle
- Node-Red Cook-Book
- SQL Server Sebenta