Relatório Final do Projeto de ETL em Pentaho Kettle

1. Enquadramento

Equipa: Bruno Tomás Lourenço Enes nº25992
Unidade Curricular: Integração de Sistemas de Informação
Curso: Engenharia de Sistemas Informáticos

---

2. Problema

Descrição:
O objetivo deste projeto é desenvolver um processo de ETL (Extract, Transform, Load) utilizando a ferramenta Pentaho Kettle, para manipular uma base de dados pertencente a uma loja online fictícia. O objetivo foi identificar e extrair clientes de nacionalidade portuguesa (código do país 351) que tenham realizado compras de produtos com um valor inferior a 200 euros. Esse processo inclui também:

- a geração em formato HTML com a lista dos clientes filtrados
- o envio via e-mail da base de dados dos clientes.

O processo final permitirá que as equipas de marketing e análise tenham acesso rápido e automatizado a dados segmentados para apoiar decisões estratégicas.

---

3. Estratégia Utilizada

Operadores e Processos Utilizados: Para atingir o objetivo do projeto, foram utilizados os seguintes operadores e processos:

- **Importação de CSVs**: Carregamento das bases de dados clientes.csv e produtos.csv.
- **Filtros e Validação com Expressão Regular**: Uso do passo Filter Rows para determinar o código do país e validação dos dígitos específicos do número de telefone por meio de expressões regulares.
- **Junção de Dados (Join Rows)**: Para associar a tabela de clientes com a de produtos com base no campo produto_compra.

- **Exportação em XML**: Adicionado um passo "XML Output" para exportar os dados resultantes do processo de filtragem para um arquivo XML. Esse arquivo XML será gerado para permitir integração com outros sistemas ou processos que precisem acessar essas informações em formato estruturado.
- **Geração de HTML**: Para criar o resultado final em formato HTML.
- **Envio de E-mails**: Para enviar o resultado diretamente aos responsáveis pelo projeto.

---

4. Transformações

Explicação das Transformações

1. **Carregamento de Dados**:
    - Utilizou-se o passo "CSV file input" para importar as tabelas clientes.csv e produtos.csv.
2. **Identificação do Código do País e Validação de Dígitos do Telefone**:
    - Para validar a autenticidade dos números de telefone, foi utilizada a expressão regular “^[0-9]{3}[1-9]{2}[0-9]{7}$” no passo "Filter Rows". Esta expressão regular permite:
        - Verificar que o número de telefone começa com três dígitos seguidos (podendo ser qualquer número de 0 a 9).
        - Assegurar que o quarto e o quinto dígitos estejam entre 1 e 9, eliminando assim números com dígitos zero nessas posições.
    - Um filtro adicional "Filter Rows" foi usado para manter apenas os clientes com telefone começando com o código "351", garantindo que apenas clientes portugueses fossem considerados.
3. **Filtro de Compras Maiores que 200 Euros**:
    - Com base no preço dos produtos, um filtro foi configurado para reter apenas as transações cujo valor do produto é superior a 200 euros.
4. **Junção de Dados**:
    - Foi utilizado o passo "Join Rows (cartesian product)" para combinar os dados de clientes e produtos com base no campo produto_compra na tabela de clientes e o campo nome na tabela de produtos.

5. **Filtragem Final**:
    - Adicionou-se um passo "Filter Rows" para manter apenas os clientes que atendem a ambas as condições:
        - O telefone passa pela validação com a expressão regular e código de país,
        - O preço do produto comprado é superior a 200 euros.

---

5. Jobs

Explicação dos Jobs

1. **generateEmail- Filtragem e Email**:
    - **Executa as Transformações**: O job carrega e executa a transformação que filtra clientes portugueses com compras acima de 200 euros.
    - **Envio de E-mail**: É enviado um email, com todos os ficheiros de output anexados, aos responsáveis pelo processo, usando o passo "Send Email".

2. **generateHTML- HTML**:
    - **Executa as Transformações**: O job carrega e executa a transformação que filtra clientes portugueses com compras acima de 200 euros.
    - **Geração de HTML**: É gerado um ficheiro HTML com o resultado das filtragens.

---

6. Vídeo com Demonstração

Abaixo está o QR Code para o vídeo demonstrativo do processo de execução DO Job de Email:

7. Conclusão e Trabalhos Futuros

Conclusão:
Este projeto mostra a versatilidade e a capacidade do Pentaho Kettle para executar processos de ETL que incluem transformações complexas, como validação por expressão regular e filtros múltiplos. O processo gerado permite segmentar automaticamente os clientes e enviar relatórios detalhados para análise.

Trabalhos Futuros:
Como melhorias futuras, podemos incluir:

- Explorar o acesso a APIs (exemplo: serviços web) remotas;
- Processos de visualização dos resultados conseguidos utilizando dashboards, se necessário com a integração de outras ferramentas como NodRed, Apache Airflow® , Home Assistant, ou outras.

---

8. Referências Bibliográficas

1. Pentaho Documentation. "Pentaho Data Integration - User Guide". Disponível em: https://help.pentaho.com .

2. BI Tools. "Transformação e Automação de Processos com Pentaho Kettle". Disponível em: https://bi-tools.com.
3. Documentação sobre o uso de CSV Input e Filter Rows no Pentaho Kettle para validação de dados com expressões regulares.