ESCOLA SUPERIOR DE TECNOLOGIA IPCA

Trabalho Prático de:
Integração de Sistemas de Informação

Realizado por:
Daniela Brito – n.º 25591

Ano Letivo 2024/2025

26 de outubro de 2024

ESCOLA SUPERIOR DE TECNOLOGIA IPCA

Índice

1. ENQUADRAMENTO

Este trabalho foi realizado no âmbito da disciplina de Integração de Sistemas de Informação (ISI) do curso de Engenharia de Sistemas Informáticos, com o objetivo de aplicar e experimentar processos de ETL (Extração, Transformação e Transferência) utilizando ferramentas adequadas, nomeadamente o Pentaho Kettle.

No cenário atual dos processos de negócio, a necessidade de integrar dados de diferentes fontes e formatos é um desafio constante. As empresas lidam com grandes volumes de dados provenientes de diversas origens e com diferentes estruturas. A capacidade de extrair, transformar e transferir esses dados de forma eficiente permite às organizações tomar decisões informadas e garantir a coerência dos seus sistemas de informação.

O objetivo deste trabalho é simular um cenário real de migração e integração de dados de clientes, com foco na conversão de dados a partir de um ficheiro CSV para formatos mais versáteis, como JSON e Excel. Este processo não só garante a limpeza e normalização dos dados, como também oferece saídas em formatos distintos, facilitando a utilização dos dados tanto em ambientes web (JSON) quanto em relatórios empresariais ou aplicações de escritório (Excel). Para completar o processo, é incluída a funcionalidade de envio automático do ficheiro Excel final por e-mail, permitindo o acesso ágil ao relatório por parte dos utilizadores finais.

O projeto proposto segue o desenvolvimento de um processo ETL com o seguinte fluxo:

1. Extrair os dados a partir de um ficheiro CSV que contém informações de clientes.
2. Transformar os dados, aplicando operações de limpeza e normalização (correção de endereços de e-mail, remoção de duplicatas).
3. Transferir os dados transformados para dois formatos distintos:
   - JSON: para utilização em sistemas web ou aplicações que consomem este formato.
   - Excel: permitindo a visualização em formato tabular e a fácil manipulação dos dados em ferramentas de escritório, como o Microsoft Excel.
   - Enviar o ficheiro Excel final por e-mail: para fornecer o resultado da transformação de dados diretamente ao utilizador, garantindo o acesso imediato ao relatório de dados processados.

Este trabalho cumpre os seguintes objetivos:

- Consolidar os conceitos associados à Integração de Sistemas de Informação.
- Aplicar ferramentas de suporte a processos ETL, nomeadamente o Pentaho Kettle.
- Realizar operações de transformação de dados, como limpeza e normalização.
- Demonstrar a importação e exportação de dados entre diferentes formatos (CSV para JSON e Excel).

- Automatizar a distribuição de resultados com o envio do ficheiro Excel final por e-mail.
- Contribuir para o desenvolvimento prático e experiência no uso de ferramentas de ETL.

2. PROBLEMA

Neste cenário prático, o problema em análise centra-se na migração e integração de dados de um ficheiro de produtos em formato CSV para formatos JSON e Excel com recurso a um processo ETL. As empresas, ao lidar com dados vindos de diferentes fontes e estruturas, frequentemente encontram dificuldades em assegurar a consistência, qualidade e adequação dos dados aos sistemas que os irão consumir. Este trabalho foca-se em resolver desafios típicos dessa migração, incluindo a limpeza, normalização e validação dos dados.

O ficheiro inicial, em formato CSV, contém dados de produtos como o *nome*, *descrição*, *preço*, *quantidade*, *e-mail do fornecedor* e *nome do fornecedor*. Porém, os dados apresentam alguns problemas que dificultam a integração noutros sistemas. Em particular, os *preços*, *e-mails* e *quantidades* encontram-se por vezes mal formatados, há entradas duplicadas e valores em falta em alguns registos. Este conjunto de dados mal estruturado compromete a precisão da informação e a sua fiabilidade para uso empresarial.

Desafios a Resolver

1. **Limpeza e Normalização de Dados**
   - **Preços**: Os preços dos produtos encontram-se em formatos variados e incorretos para o sistema, como o uso de ponto em vez de vírgula para separar as casas decimais, ou até mesmo valores não numéricos. O processo ETL inclui a transformação para um formato coerente de preços (ex.: NN,NN€), garantindo que todos os valores estão prontos para futuras operações financeiras e análises.
   - **E-mails**: Os endereços de e-mail dos fornecedores apresentam erros como o uso de caracteres incorretos e falta de estrutura válida (ex.: falta do "@" ou domínio incorreto). A validação e normalização dos e-mails permitirá a integração correta com outras aplicações e a criação de canais de contacto consistentes com os fornecedores.

2. **Remoção de Registos Duplicados**

   Alguns produtos, como "Cadeira" e "Mesa", aparecem duplicados, contendo os mesmos valores em todas as suas colunas. O sistema ETL é configurado para identificar e remover estas entradas repetidas, assegurando que cada produto tem apenas uma linha no relatório final, evitando redundâncias.

3. **Tratamento de Dados Incompletos**

   Alguns registos estão incompletos, como é o caso de produtos com a quantidade em branco. Estes dados serão tratados de forma a serem identificados para correção ou verificação adicional, ou marcados para revisão.

Objetivo do Processo ETL

Para além das etapas de transformação, este trabalho integra a funcionalidade de envio automático do ficheiro Excel final por e-mail. Esta solução oferece ao utilizador final um acesso direto e imediato ao relatório de produtos em formato Excel, uma prática comum no contexto empresarial onde a partilha de relatórios é fundamental. Assim, o trabalho pretende demonstrar a execução de cada uma das fases do processo ETL, com especial atenção à correção e consistência dos dados e à sua apresentação final em diferentes formatos, possibilitando a integração dos dados em sistemas web e de análise empresarial.

Para atingir estes objetivos, é utilizada a plataforma Pentaho Kettle, uma ferramenta de ETL, que facilita a automação de processos, assegura a qualidade dos dados e permite a interoperabilidade entre formatos distintos, essencial em qualquer projeto de Integração de Sistemas de Informação.

3. ESTRATÉGIA UTILIZADA

Para a resolução do problema de **integração e transformação de dados de produtos**, adotou-se uma abordagem estruturada de ETL (Extração, Transformação e Transferência) através da ferramenta *Pentaho Kettle*. A estratégia foca-se na extração de dados de um ficheiro CSV, transformação com correções e validações, e transferência dos dados para formatos JSON e Excel, garantindo a sua adequação a diferentes contextos de uso empresarial. Este processo também inclui o envio do relatório final em Excel para um e-mail específico, facilitando a distribuição e consulta dos dados.

Operadores e Processos Envolvidos

1. **Extração dos Dados (Extração)**

   A primeira etapa do processo ETL consiste na leitura do ficheiro CSV, que contém campos como *Produto*, *Descrição*, *Preço*, *Quantidade, Email_Fornecedor* e *Fornecedor*. A ferramenta Kettle utiliza um operador de entrada que permite definir o ponto-e-vírgula como delimitador dos campos. Esta configuração permite interpretar corretamente os valores monetários em euros, onde a vírgula separa as casas decimais (ex.: `159,99€`).

2. **Transformação dos Dados (Transformação)**

   a) **Limpeza de Dados**: Esta etapa inclui operações de verificação e correção para garantir a consistência dos dados:
   - **Remoção de registos duplicados**: Identificação e eliminação de registos repetidos (ex.: duplicações de produtos como "Cadeira" e "Mesa") através do step *Unique Rows*, para garantir que cada produto está representado de forma única no ficheiro final.
   - **Preenchimento de campos vazios**: Identificação de valores ausentes (ex.: campos de quantidade em branco) para preenchimento ou marcação como inválidos, de forma a manter a integridade dos dados.

   b) **Normalização e Correção de Preços:**

   Para os preços, são aplicadas transformações que substituem pontos por vírgulas e garantem o formato correto, incluindo as duas casas decimais (ex.: transformar `29.99` em `29,99€`). Esta etapa assegura que os dados estejam consistentes para processamento e análise.

   Exemplo de expressão regular:

   ^\d{1,3}(,\d{2})?€

   - **^**: Início da linha.
   - **\d{1,3}**: Entre 1 a 3 dígitos antes da vírgula.

- **(,\d{2})?**: Uma vírgula seguida de dois dígitos (decimal opcional).
- **€**: Símbolo do euro.
- **$**: Fim da linha.

c) Validação de E-mails:

A correção dos e-mails de fornecedores utiliza expressões regulares (ER) que garantem que todos os endereços estão no formato padrão (ex.: `nome@dominio.com`). Esta validação abrange:

- Início e fim da string: Utilização de símbolos `^` e `$` para assegurar que o e-mail corresponde exatamente ao formato.
- Verificação de caracteres permitidos antes do @: Apenas letras, números e os caracteres `.`, `_`, `%`, `+` e `-`.
- Formato de domínio: Garantir que o e-mail inclui um domínio válido após o `@`, com uma extensão mínima de dois caracteres, como `.com` ou `.pt`.

Exemplos como `fornecedor8@exemplo,com` (com uma vírgula em vez de ponto) e `fornecedor10@exemplo` (domínio incompleto) são corrigidos ou identificados para revisão, garantindo a adequação dos contactos de fornecedores.

Expressão Regular:

^[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}$

- **^**: Início da linha.
- **[a-zA-Z0-9._%+-]+**: Permite letras, números e certos caracteres especiais antes do @.
- **@**: Símbolo obrigatório.
- **[a-zA-Z0-9.-]+**: Letras, números, pontos e hífens no domínio.
- **\.**: Ponto literal.
- **[a-zA-Z]{2,}**: Extensão do domínio com pelo menos 2 letras.
- **$**: Fim da linha.

3. Transferência dos Dados (Transferência)

a) A última etapa consiste na **exportação dos dados transformados** para dois formatos distintos:

- **JSON**: Um formato leve e amplamente usado em aplicações web e APIs. A exportação para JSON permite que os dados sejam facilmente integrados em sistemas web e reutilizados em ambientes de desenvolvimento digital.
- **Excel**: A exportação para Excel permite que os dados sejam visualizados em formato tabular e manipulados por aplicações de escritório, como o Microsoft Excel, facilitando a análise e preparação de relatórios.

b) Envio do Relatório Final por E-mail:

Para automatizar a distribuição dos dados, o ficheiro Excel final é enviado por e-mail a um destinatário específico, utilizando um Job que inclui a configuração do servidor de e-mail, destinatário e anexos. Esta etapa é essencial para garantir a rápida disponibilização dos dados tratados aos utilizadores finais, diretamente no seu ambiente de trabalho.

Síntese dos Objetivos da Estratégia Utilizada

Com esta abordagem, demonstra-se a aplicabilidade do ETL para resolver problemas de integração de dados em diferentes formatos e assegurar a consistência e qualidade das informações. Este fluxo de trabalho simula um cenário empresarial real, onde dados provenientes de um ficheiro CSV são convertidos, validados e distribuídos automaticamente, oferecendo uma solução prática e eficiente para a gestão e integração de sistemas de informação.

4. TRANSFORMAÇÕES

5. JOBS

6. VÍDEO (QR CODE)

7. CONCLUSÃO

8. BIBLIOGRAFIA

O projeto