Relatório de Trabalho Prático

"Tratamentos de dados com recurso a processos ETL, utilizando Pentaho Kettle"

Carlos Daniel Melo Miranda Oliveira
N° 18835 - Regime: Diurno

Docente: Luís Ferreira
U.C.: Integração dos Sistemas de Informação
Ano letivo 2020/2021

Licenciatura em Engenharia de Sistemas Informáticos
Escola Superior de Tecnologia
Instituto Politécnico do Cávado e do Ave

Resumo

A gerência de uma empresa é uma atividade que requer ter bastante conhecimento e e poder de decisão. Para expandir o negócio e melhorar os processos deste, um gerente precisa de conseguir acompanhar as novidades do mercado e adotar estratégias e recursos eficientes na rotina do trabalho de modo a tornar o seu negócio mais produtivo e com qualidade igual ou superior.
Uma das ferramentas que melhora e facilita na execução de processos numa empressa é o Pentaho. O Pentaho é uma ferramenta muito poderosa de código aberto que possibilita organizar, analisar de forma rápida e precisa uma grande quantidade de dados, ajuda na exploração de dados (data mining), workflow e capacidades de ETL — Extração, Tratamento e Limpeza de dados.
As soluções em Business Intelligence estão a conquistar cada vez mais espaço no mercado e é extremamente recomendado utilizá-las visto que facilitam bastante os processos e permitem obter resultados positivos de uma maneira mais rápida.

Palavras-Chave: Pentaho, Business Intelligence, ETL, dados

Conteúdo

Lista de Figuras

Lista de Tabelas

Siglas

HTML hypertext markup language

CSV Comma-separated values

XML extensible markup language
XSL XML Style Language

Símbolos

π pi

$\mathcal{S}$ a set

$\mathcal{U}$ universal set

Capítulo 1

Introdução

No capítulo introdutório será discutido o contexto do problema. De seguida é apresentada a motivação e objetivos do projeto e por fim é descrita a estrutura do documento.

1.1 Contextualização

Para muitas entidades/empresas o Excel continua a ser uma ferramenta com bastante importância, isto devido à sua facilidade e intuitividade de uso e porque é possível, através dele, tirar partido de ferramentas muito poderosas. Contudo a sua visualização poderá ser mais massiva no caso de querer retirar alguma informação mais específica e existe um maior risco de perda da mesma.

1.2 Motivação e objetivos

O tema deste projeto baseia-se em, tirando partido da ferramenta Pentaho Kettle, tratar dados com recurso a processos ETL. Ou seja, extrair, transformar e carregar dados. Para o efeito do trabalho serão usados dados de crimes efetuados em Portugal por região, assim como dados com concelhos e distritos de um repositrio "Central de Dados".

1.3 Estrutura do Documento

O documento encontra-se organizado em 3 capítulos, detalhados de seguida. O capítulo Introdutório onde é detalho o projeto e feita uma abordagem ao contexto do problema, motivação e objetivos. O capítulo de Implementação, onde é descrito o problema e como o projeto foi implementado na vertente de programação, de modo a alcançar os objetivos, bem como os resultados obtidos. Por fim, o capítulo de conclusão, onde são retiradas as conclusões do desenvolvimento da ferramenta.

Capítulo 2

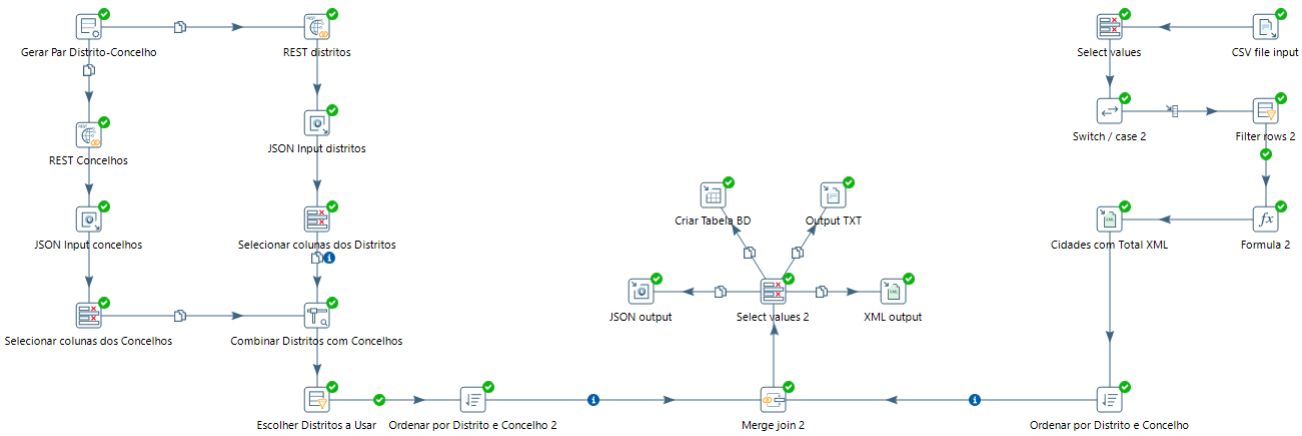
Implementação

No capítulo de Implementação é efetuada uma descrição do problema bem como a solução encontrada para o mesmo, assim como algumas tentativas de implementação que nao foram conseguidas, descrevendo sucintamente todos os passos tomados na resolução.

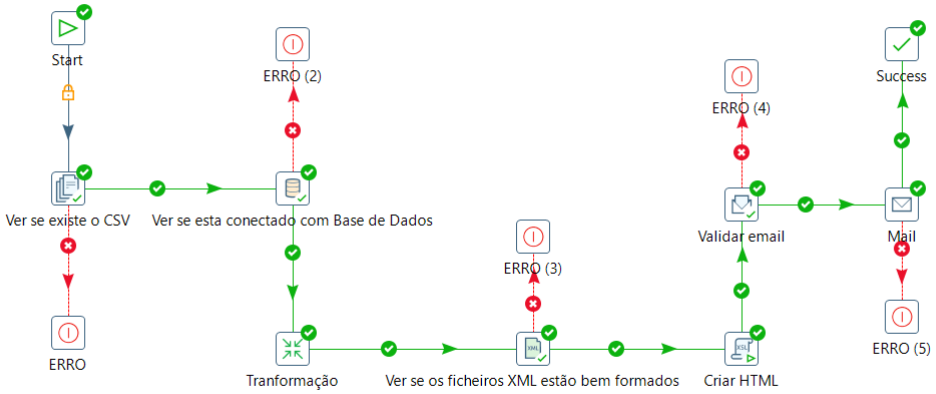
2.1 Descrição do problema

Tendo em conta a quantidade de crimes que ocorreram ao longos dos anos até aos dias de hoje pareceu pertinente organizar todos os dados obtidos dos crimes desde 1993 até ao ano de 2014, em diversas regiões de Portugal, e realizar várias transformações sobre o ficheiro inicial de modo a obter algo de consulta mais fácil e maior segurança em guardar os dados assim como completar com alguma informação que pudesse faltar acerca dos concelhos e distritos. Este documento contém informação não muito relevante assim como uma visualização bastante complexa, o que impossibilita a extração de informação de forma rápida e intuitiva.

2.2 Solução

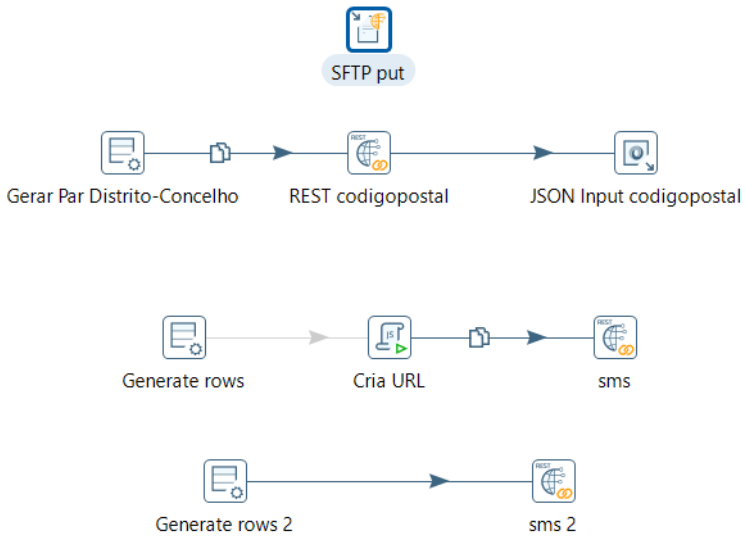


A solução consiste em tirar partido do Pentaho Kettle e a partir de várias transformações conseguir atingir o objetivo. Para isso comecei por fazer o input do "Crimes input CSV"e retirar as colunas de dados que não tenham informação relevante, a partir daqui apenas escolhi 3 distritos de forma a exemplificar tratamentos de dados, fiz isto recorrendo a um "Switch/Case"escolhendo a coluna que queria filtrar. Apliquei um filtro onde escolhi apenas Aveiro, Braga e Beja, agora com apenas estes dados apliquei uma "Formula"ao qual pude acrescentar uma nova coluna (Total19932014crimes) somando todas as colunas dos anos, e passei toda esta informação para um ficheiro em formato XML nomeado "Cidades com Total XML". Agora com este ficheiro preparado a nível de informação fiz uma ordenação ascendente pelas colunas dos distritos e dos concelhos e passei para o tratamento de outros dados. Tendo isto em conta dirigi-me ao repositório "Central de Dados"onde fui buscar o URL de 3 ficheiros: "concelhos.json", "distritos.json", "codigospostais.json", usei um "Generate Rows"para criar estas 3 novas linhas e usei um "Rest Cliente"para extrair os dados do URL configurei o "Rest Client"para aceitar um URL do campo e coloquei o método do HTTP como GET e usei "json input"para pôr os dados de cada URL.
Por consequente selecionei apenas as colunas que queria trabalhar de ambos os ficheiros e usei o "Stream Lookup"para conseguir juntar a informação dos dois ficheiros, visto que ambos tinham o "codigodistrito"em comum. Posto isto usei novamente um filtro para escolher os mesmos distritos que tinha escolhido anteriormente (Aveiro, Braga, Beja), mas desta vez tentei restringir apenas com condições REGEX e ordenei também os dados ascendentemente por distrito e concelho.
Agora com toda a informação necessária ordenada tentei fazer "Merge Join"para conseguir por toda a informação unida, recorrendo a várias tentativas para perceber a funcionalidade da transformação consegui executar, usando o "full outer"como tipo de junção. Como fiquei com informação repetida, retirei agora as colunas que nao interessavam e para finalizar exportei a informação para 3 tipos de formato diferente: json, xml e txt. Para concluir a transformação resolvi experimentar o output como tabela, criei entao uma base de dados PostGres, fiz a conexão à mesma e consegui executar o código SQL para a mesma, criando assim uma tabela "Crimes"com toda a informação gerada. Dada por concluída a transformação parti para a criação de um "Job"este inicia e verifica logo a existência do meu ficheiro CSV de input usado na transformação, caso nao exista é cancelada a operação. Seguidamente é verificada a conexão à base de dados usada na trasnformação que, por sua vez, se não existir também é cancelada a operação.
Imediatamente segue-se a transformação em si e mal esta seja completada é verificado se os ficheiros XML gerados estão bem formados, não o sendo é cancelada a operação. Neste momento é criado um HTML com o XML final e um XSL feito para o intuito. Para finalizar iremos enviar uma informação de que o processo foi executado com sucesso para um email, este é validado antes e se for válido procede para o envio. Se alguma destas operações finais falhar também é cancelada a operação.



2.3 Tentativas de Implementações

Houve tentativas de implementação que falharam, no Pentaho tem uma transformação, apenas com essas tentativas, representadas de uma maneira simples e individual. Uma delas foi o uso de um servidor FTP com o programa FileZilla, no Pentaho era posto um "SFTP put"onde tentei configurar o diretório remoto e o local. Para alem disto, também foi tentado usar um "Rest Client"com dados de códigos postais, contudo a operação estava a demorar bastante tempo a carregar todos os dados, pelo que não usei no projeto. Por fim foi também experimentado a possibilidade de envio de SMS via telemóvel com recurso de uma api própria, mas sem sucesso, uma api tentada foi a "localsms"mas nao permitia uso de números portugueses, outra foi a api "SMS BULK"a qual tinha mais esperança mas mesmo assim nao consegui executar. Estas tentativas poderão ser exploradas mais e até serem usadas em trabalhos futuros, visto que me pareceram bastante interessantes.



Capítulo 3

Conclusão

Neste capítulo final é abordado o problema como um todo, bem como a solução final obtida.

3.1 Lições aprendidas

O desenvolvimento deste projeto permitiu-nos enquanto alunos colocar em prática o conteúdo lecionado até a data na U.C. assim como explorar e investigar mais acerca da ferramenta Pentaho Kettle, e perceber que realmente é uma ferramenta muito poderosa e que à base de simples "cliques"conseguimos efetuar operações com alguma complexidade a nível de programação. Assim como interligar diferentes tipos de ferramentas, como base de dados, ficheiros externos, como o caso do repositório "Central Dados"de uma maneira muito intuitiva.

3.2 Apreciação final

Numa última nota é de salientar que o desenvolvimento deste projeto foi concluído com sucesso, tendo a aplicação mostrado ótimos resultados durante a fase de testes. Numa primeira fase do projeto foi definido o tema a tratar e que dados usar para poder trabalhar. Depois foi feito um estudo relativamente a diferentes operações que poderiam ser feitas e fossem de certa forma relevantes ao trabalho, por fim foi tentar implementar essas operações de forma a transformar os dados da melhor maneira. No final, após testes e correções nos processos de ETL, foi traçada uma análise ao projeto como um todo, tendo correspondido aos objetivos propostos.

Bibliografia

[1] H. Partl: *German TEX*, TUGboat Volume 9, Issue 1 (1988)

[2] Ferreira, L.(2020) Diapositivos e Sebentas da Unidade Curricular Integração dos Sistemas da Informação da Licenciatura em Engenharia de Sistemas Informáticos. Barcelos, Braga, Portugal.