Construção de Processo Automatizado de ETL para Acompanhamento de Registro de Marcas no INPI

Christian Testtzlaffe Alpoim[1]

[1]Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) – Rio de Janeiro – RJ – Brasil

chrufes@yahoo.com.br

Abstract. *This paper describes the development of an automated flow of extraction, transformation and loading (ETL) of trademark registration processes in Brazil, with Python, SQL Power Architect, Pentaho Data Integration and PostgreSQL.*

Resumo. *Este artigo descreve o desenvolvimento de um fluxo automatizado de extração, transformação e carga (ETL) de processos de registro de marca no Brasil, com Python, SQL Power Architect, Pentaho Data Integration e PostgreSQL.*

1. Objetivo

Este trabalho tem o objetivo de construir um fluxo automatizado de coleta, transformação e carga/persistência de dados dos processos de registros de marcas originados pelo Instituto Nacional de Propriedade Industrial (INPI).

Antes deste processo de extração, transformação e carga (ETL), é necessário o mapeamento dos dados, a modelagem e a criação do repositório destes registros (banco de dados relacional).

Uma vez que o processo esteja configurado e sendo executado periodicamente, este banco de dados permitirá a construção de aplicações para acompanhamento e alerta sobre o avanço nos registros de marcas.

2. INPI

O Instituto Nacional da Propriedade Industrial (INPI) é uma autarquia federal brasileira, criada em 1970, vinculada ao Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC).

Conforme art. 2º da Lei nº 5.648/1970, "O INPI tem por finalidade principal executar, no âmbito nacional, as normas que regulam a propriedade industrial, tendo em vista a sua função social, econômica, jurídica e técnica, bem como pronunciar-se quanto à conveniência de assinatura, ratificação e denúncia de convenções, tratados, convênios e acordos sobre propriedade industrial". (Redação dada pela Lei nº 9.279, de 1996).

Entre suas funções, o INPI é responsável pelo registro e concessão de marcas, que é a base para este trabalho.

3. Etapas, Arquitetura e Tecnologias

O diagrama abaixo resume as etapas do processo, bem como a arquitetura da solução e a indicação das tecnologias.

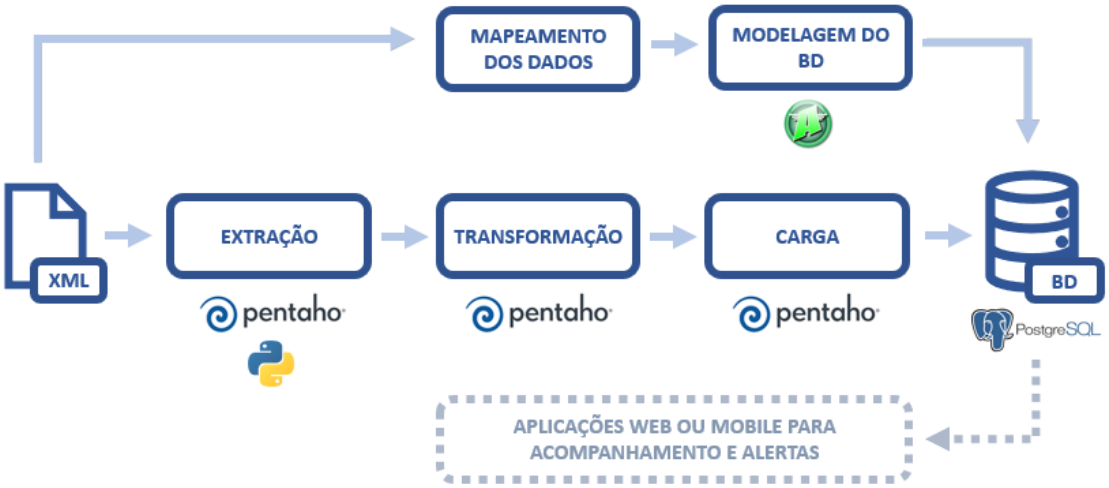

Figura 1. Etapas, Arquitetura e Tecnologias

Fonte dos dados: arquivos XML do site do INPI.

Mapeamento dos dados: estudo das principais tags e atributos do arquivo, bem como suas relações.

Modelagem dos dados: desenho do modelo de tabelas e seus relacionamentos, com uso do SQL Power Architect. Por meio do Power Architect, é relativamente simples modelar as entidades e suas relações, sendo uma ferramenta bastante utilizada por profissionais especialistas em banco de dados. Além disso, tem integração com diversos sistemas gerenciadores de BD, inclusive o PostgreSQL, utilizado neste trabalho.

BD (banco de dados): criação e atualização das tabelas por meio do PostgreSQL. O PostgreSQL foi definido como o sistema gerenciador de banco de dados do projeto, por ser de código aberto, com mais de 30 anos de desenvolvimento. Tem uma boa reputação e é amplamente utilizado pelo mercado, com arquitetura comprovada, confiável e com recursos robustos. Além disso, pode ser executado em todos os principais sistemas operacionais.

ETL: extrações, transformações e cargas com Pentaho Data Integration (PDI). O PDI tem inúmeras funcionalidade de ETL que facilitam a coleta, a limpeza, as transformações e a persistência dos dados. É intuitivo na utilização, e tem integração com vários formatos de entrada e com diversos sistemas gerenciadores de banco de dados, inclusive PostgreSQL. Permite agendar processos para serem executados automaticamente, bem como gerar alertas de finalização ou falhas. O Pentaho Data Integration é amplamente utilizado por diversos tipos de clientes, como instituições financeiras, indústrias, órgãos dos governos federal, estaduais e prefeituras, entidades de saúde, universidades, entre outros. Para a primeira tarefa do ETL, o download dos

arquivos .zip, foi utilizado um script em Python. Python é uma linguagem de programação de alto nível, lançada em 1991. É de propósito geral, sendo muito utilizada para Ciência de Dados e scripts.

4. Fonte dos Dados

Os dados a serem coletados para o processo são públicos e disponibilizados na Revista da Propriedade Industrial (RPI). O INPI divulga semanalmente, toda terça-feira, as atualizações sobre os registros de marcas, tanto no formato PDF quanto em XML (compactado/zipado).

Para este trabalho, foi avaliado que o arquivo XML contém as principais informações de despachos do INPI, e servirá como base para o processo de ETL. Vale destacar que este arquivo, quando descompactado, tem entre 20 e 30MB, com dezenas de milhares de despachos.

5. Mapeamento dos Dados

Inicialmente, foi necessário o estudo dos dados de registro de marcas no INPI. Foram avaliadas minuciosamente algumas Revistas de Propriedade Industrial, tanto na versão XML quanto em PDF. No XML, há diversos nós (tags) e atributos, alguns sempre presentes, e outros aparecem a depender do tipo de despacho.

Como exemplo, há a tag "titular". Nela estão presentes os atributos nome, país e UF do titular da marca. A UF é preenchida apenas quando o país é BR (Brasil).

Tabela 1. Tags e Atributos do Arquivo XML

Tag	Caminho	Atributos
revista	/revista	numero, data
processo	/revista/processo	numero, data-deposito, data-concessao, data-vigencia
despachos	/revista/processo/despachos	
despacho	/revista/processo/despachos/despacho	codigo, nome
texto-complementar	/revista/processo/despachos/despacho/texto-complementar	
texto-sobrestamento	/revista/processo/despachos/despacho/texto-sobrestamento	
protocolo	/revista/processo/despachos/despacho/protocolo	numero, data, codigoServico
requerente	/revista/processo/despachos/despacho/protocolo/requerente	nome-razao-social, pais, uf
procurador	/revista/processo/despachos/despacho/protocolo/procurador	
cedentes	/revista/processo/despachos/despacho/proto	

	colo/cedentes	
cedente	/revista/processo/despachos/despacho/protocolo/cedentes/cedente	nome-razao-social, pais, uf
cessionarios	/revista/processo/despachos/despacho/protocolo/cessionários	
cessionario	/revista/processo/despachos/despacho/protocolo/cessionarios/cessionário	nome-razao-social
marca	/revista/processo/marca	apresentacao, natureza
nome	/revista/processo/marca/nome	
titulares	/revista/processo/titulares	
titular	/revista/processo/titulares/titular	nome-razao-social, pais, uf
procurador	/revista/processo/procurador	
sobrestadores	/revista/processo/sobrestadores	
sobrestador	/revista/processo/sobrestadores/sobrestador	processo, marca
lista-classe-nice	/revista/processo/lista-classe-nice	
classe-nice	/revista/processo/lista-classe-nice/classe-nice	codigo
especificacao	/revista/processo/lista-classe-nice/classe-nice/especificação	
traducao-especificacao	/revista/processo/lista-classe-nice/classe-nice/traducao-especificacao	
status	/revista/processo/lista-classe-nice/classe-nice/status	
classes-vienna	/revista/processo/classes-vienna	
classe-vienna	/revista/processo/classes-vienna/classe-vienna	codigo, edicao
classe-nacional	/revista/processo/classe-nacional	codigo
especificacao	/revista/processo/classe-nacional/especificação	
sub-classes-nacional	/revista/processo/classe-nacional/sub-classes-nacional	
sub-classe-nacional	/revista/processo/classe-nacional/sub-classes-nacional/sub-classe-nacional	codigo
dados-de-madri	/revista/processo/dados-de-madri	numero-inscricao-internacional, data-recebimento-inpi

apostila	/revista/processo/apostila	
prioridade-unionista	/revista/processo/prioridade-unionista	
prioridade	/revista/processo/prioridade-unionista/prioridade	data, numero, pais

6. Modelagem das Tabelas e dos Campos

Partindo do mapeamento dos dados da etapa anterior, foi possível identificar quais as entidades mais importantes para o modelo do banco de dados. Basicamente, o escopo envolve: revista, processo, marca, titular, procurador, despacho, classe-nice e classe-vienna.

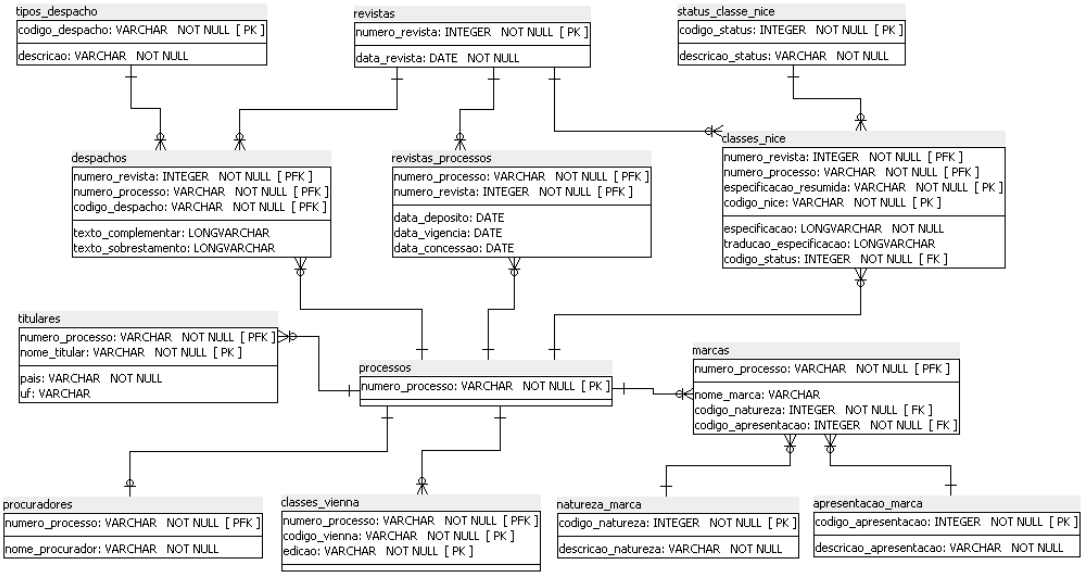

Figura 2. Modelo Relacional

7. Criação do Banco de Dados

O trabalho utiliza como Sistema Gerenciador de Banco de Dados (SGBD) o PostgreSQL. A criação das tabelas foi originada diretamente do SQL Power Architect, tomando como base o modelo desenhado.

8. Extração, Transformação e Carga dos Dados

O processo central do trabalho, o ETL, segue algumas etapas importantes para o alcance do objetivo de geração do banco de dados devidamente carregado.

Extração: etapa de coleta / extração dos recursos de dados. Neste trabalho, a extração é responsável pelo download, descompactação dos arquivos zipados, leitura do XML e loop sobre as tags importantes para obtenção dos dados.

Transformação: tratamento dos dados brutos, formatações e adequações do conteúdo dos campos. Aqui foram realizadas operações sobre textos (strings), formatações de datas, seleção de valores, mapper, entre outros.

Carga: etapa final do ETL, persiste os dados no repositório de destino. No caso deste projeto, o conjunto de dados extraídos e transformados para cada entidade são carregados para o PostgreSQL, de forma que semanalmente os registros de processos de marcas estejam sempre atualizados.

Para que seja executado sem necessidade de ação manual, foi configurado no "Agendador de Tarefas" do Windows, para toda terça-feira, duas execuções: 1) download do arquivo .zip do site do INPI; 2) job configurado no Pentaho Data Integration (executado em background).

Para o download, foi desenvolvido um script em Python, que verifica qual o número atual da revista e baixa o arquivo para determinada pasta.

No Pentaho Data Integration (PDI), foi implementado um fluxo (job) com etapas de extração dos dados, transformações / formatações / preparação dos campos, e carga para cada tabela no PostgreSQL, conforme telas do PDI:

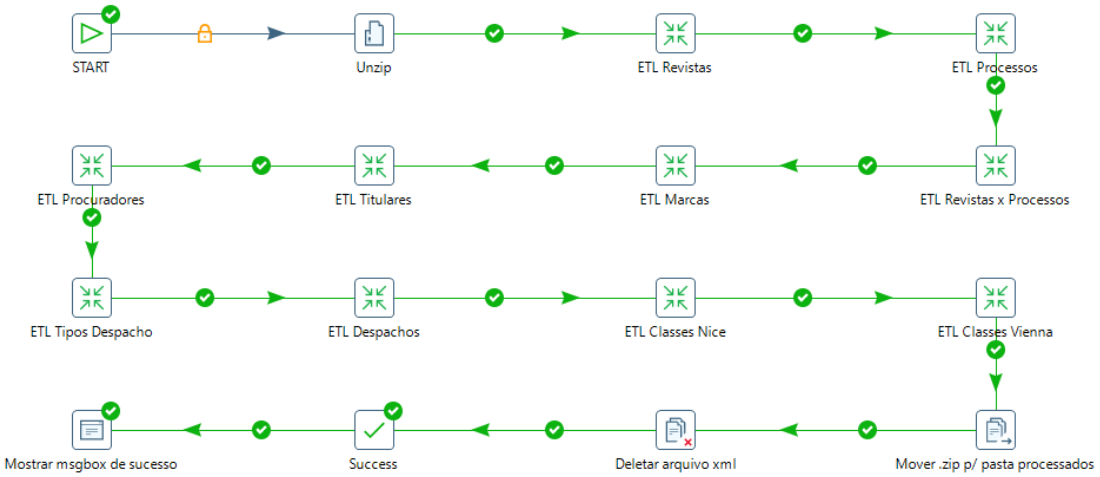

Figura 3. Jobs no PDI

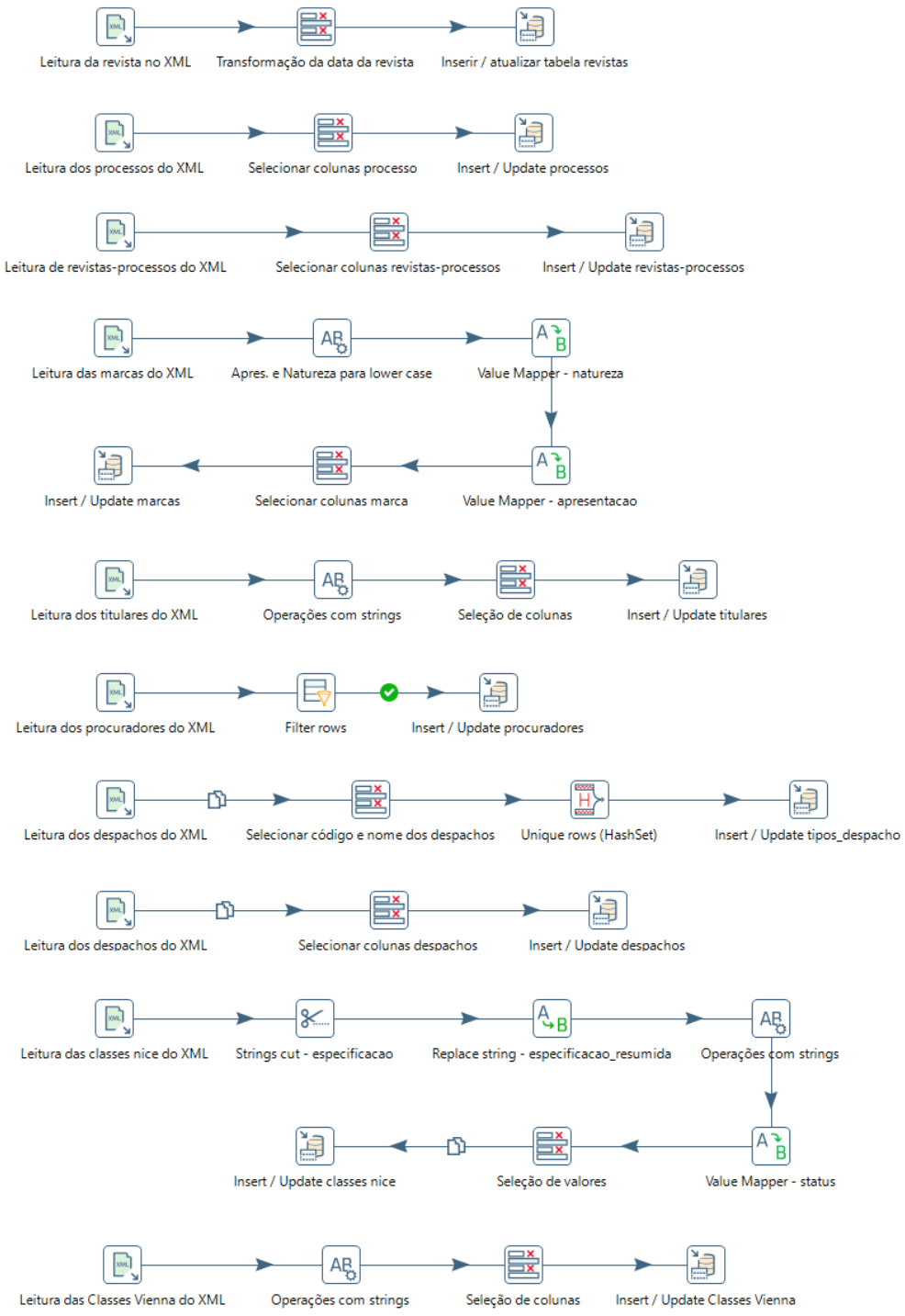

Figura 4. Transformations no PDI

A primeira tarefa do job é "unzip", que extrai o XML do arquivo compactado para o diretório definido. Segue com o job organizando uma sequência de chamadas às "transformations". Cada uma realiza a leitura dos dados importantes do XML, efetua as transformações de alguns campos e preenche as respectivas tabelas no banco de dados. Por fim, move os arquivos compactados para uma pasta auxiliar de histórico e deleta o XML utilizado, limpando a pasta para a carga da semana seguinte.

8. Resultados e Conclusões

A sequência de etapas acima atingiu o objetivo do trabalho de construir um fluxo automatizado, que semanalmente realiza o ETL e que culmina na persistência dos dados no PostgreSQL.

As tecnologias adotadas foram capazes de garantir esta carga. O SQL Power Architect se mostrou intuitivo e efetivo na modelagem das entidades e na criação das tabelas. O Python atendeu perfeitamente a etapa de download com um código enxuto. O Pentaho Data Integration foi versátil, permitindo diversos tipos de transformações, e relativamente rápido na leitura de arquivos XML de mais de 20 MB, e na carga direta no PostgreSQL. Neste fluxo semanal, uma Revista de Propriedade Industrial é lida e carregada no banco de dados entre 1 e 2 minutos. E o PostgreSQL, conforme esperado, se mostrou robusto para cadastro de milhões de registros e com boa latência nas consultas.

No trabalho, foram carregadas as últimas 53 Revistas, equivalente a um ano de atualizações do INPI. Entre os números do trabalho, foram persistidos mais de 1 milhão de despachos, referentes a mais de 700 mil processos de registro de marca. O banco de dados final contém mais de 1 GB.

Este sucesso no resultado permite que o projeto evolua por exemplo para soluções web ou mobile a fim de disponibilizar serviços de acompanhamento de registros de marca no INPI e de alertas sobre atualizações de despachos, com base nos dados gravados.