Logotipo

Descrição gerada automaticamente

DIRETORIA DE EDUCAÇÃO CONTINUADA - IEC

Disciplina: **ARQUITETURAS E SERVIÇOS DE DATA LAKES E DATA WAREHOUSING**

Nome do curso: **ENGENHARIA DE DADOS**

Professor: Ricardo Brito Alves

Trabalho Prático – Criar uma transformação de dados com o Pentaho PDI.

Trabalho em grupo.

# Pentaho – Transformação de dados

Objetivos

Exercitar os seguintes conceitos:

* ETL.
* Ferramentas ETL – Pentaho PDI e Apache Hop.
* Processos de carga.
* Orquestração.

Enunciado

Vamos utilizar dois arquivos de entrada com informações da Capes onde é feito uma avaliação da Pós-Graduação Stricto Sensu. A avaliação da pós-graduação stricto sensu tem como objetivo a certificação da qualidade da pós-graduação brasileira bem como a identificação de assimetrias regionais e de áreas estratégicas do conhecimento. Ela é orientada pela Diretoria de Avaliação/CAPES e realizada com a participação da comunidade acadêmico-científica por meio de consultores ad hoc. Tem como pilares a formação pós-graduada de docentes para todos os níveis de ensino e de recursos humanos qualificados para o mercado, bem como o fortalecimento das bases científica, tecnológica e de inovação.

Estes dois arquivos servirão como entradas de dados para executarmos algumas atividades de transformação no Pentaho DTI.

Entrega

Como entrega você precisa enviar um arquivo compactado com os arquivos de transformação, arquivos com extensão “.ktr”.

Atividades

Basicamente iremos construir uma transformação que ao final irá alimentar uma tabela Dimensão chamada **Dim-Produção**.

**Prática Pentaho PDI**

Interface gráfica do usuário

Descrição gerada automaticamente com confiança média

**Atividades**

Processo de carga:

O primeiro passo é baixar os arquivos o link abaixo ou pegá-los na plataforma do IGTI:

<https://drive.google.com/drive/u/0/folders/1wpAisTfHDWJGnhVi9S0tMux1aUpDt2RW>

*- Producao-2018-Bibliografica-Artigo.xls.*

*- Programa-2018-Capes.csv.*

Para iniciar, abra o **Pentaho PDI (Spoon.bat).**

Vamos criar o repositório apontando para o diretório do nosso projeto:

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Depois vamos criar uma transformação, um arquivo .ktr:

(*menu File | Novo | Transformação*)

Abra a categoria input e selecione os **Steps** ***Excel Input*** *e* ***CSV file Input****.*

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Logotipo

Descrição gerada automaticamente com confiança média

**Edite o Step Microsoft Excel input (grupo input no Design) com os seguintes parâmetros:**

**Aba Files**

* **File or directory:** informe o arquivo Producao-2018-Bibliografica-Artigo.xls. Para ter certeza de que o arquivo foi localizado clique no botão show filename para certificar que ele está sendo exibido.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

* **Aba Sheets**
* Clique no botão Get sheetname e escolha a planilha desejada.

Interface gráfica do usuário, Aplicativo, Tabela

Descrição gerada automaticamente

**Aba Content**

* Certifique-se que o campo Header esteja marcado.
* Encoding: para aceitar a acentuação sem mostrar como caracteres especiais, em geral, o mais utilizado seria o UTF-8, visto que suporta todo o alfabeto e acentos além de uma gama gigantesca de caracteres especiais, mas isso pode variar e vamos usar o ISO-8859-1.
  + Mude o campo **Encoding** para **ISO-8859-1**.

**Aba Fields**

* Clique no botão Get fields from header now e veja todos os campos disponíveis.
* Dê uma olhada nos dados que serão extraídos do arquivo clicando no botão Preview rows. Clique Ok e salve a transformação.

**Edite o Step CSV file input com os seguintes parâmetros:**

* Filename: localize o arquivo Programa-2018-Capes.csv com o botão Navegar.
* Delimiter: informe: “;” (ponto e vírgula).
* Mude o campo Encoding para ISO-8859-1.
* Desmarque a opção Lazy conversion.
* Clique no botão Obtem campos e veja os campos que serão lidos.
* Clique no botão Preview para visualizar uma amostra dos dados.
* Retire o símbolo da moeda (R$) da propriedade Currency. Basta apagar essa informação em todos os atributos.
* Confirme se o Header row present está marcado. Isso indica que o arquivo tem um header com o nome dos campos.

**Adicione o Step Stream lookup e faça as ligações conforme figura abaixo.**

Diagrama

Descrição gerada automaticamente com confiança baixa

**Selecione o Step Stream lookup com os seguintes parâmetros:**

* Lookup Step: escolha o Step do CSV.
* Clique nos botões Get fields e Get lookup fields. O lookup fará a ligação dos arquivos pela chave comum. É uma ligação de 1 para N, ou seja, um Programa tem N Publicações e a ligação entre os arquivos se dá pelo campo CD\_PROGRAMA\_IES. Temos esse campo nos dois arquivos e é o identificador do Programa.

Tabela

Descrição gerada automaticamente

Na grade de cima, remova todos os campos deixando apenas o campo CD\_PROGRAMA\_IES.

Na grade de baixo deixe os campos que deseja que sejam mostrados a partir da junção.

Clique com o botão direito em cima do Step Stream lookup e escolha as opções abaixo de visualização.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Suponha que não precisamos de todos os campos vindos da junção e para isso é necessário inserir o Step Select values fazendo a ligação necessária.

Uma imagem contendo Mapa

Descrição gerada automaticamente

**Aba Select & After**

* mostra os campos após você clicar no botão Get fields to select.

**Aba Remove**

* mostra os campos após você clicar no botão Get fields to remove. Os campos que permanecerem nessa lista serão removidos.

**Ainda no Step Select values**

* você vai renomear o campo CD\_PROGRAMA\_IES para CODIGO-IES.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

**Insira o Step Text file output com os seguintes parâmetros:**

**Aba File**

* **Filename**: caminho onde irá salvar o arquivo + nome do arquivo txt.

**Aba Fields**

* Clique no botão Obtém campos e veja os campos que serão gravados.
* Você pode fazer alterações diretamente na grade. Clique no botão **Minimal width** e veja que o Step fornece um formato padrão para os campos.

**Rode** o arquivo de transformação e veja se gravou o arquivo texto no diretório indicado.

Observe que inicialmente dará um erro referente a linha 878. Na realidade teremos erro em duas linhas. Isso porque há uma tentativa de converter valor no campo CD\_CONCEITO\_PROGRAMA de String para Integer.

Texto

Descrição gerada automaticamente com confiança baixa

Você deve tratar o erro na transformação para que o processo não seja interrompido.

Insira o **Step Dummy (do nothing)** e ligue o **Step CSV** ao **Step Dummy**. Na ligação selecione **Error handling for Step**. Aparecerá uma caixa onde você deve aceitar a distribuição pelo botão Distribuir.

Uma imagem contendo Diagrama

Descrição gerada automaticamente

Rode novamente e veja que será concluído com sucesso.

Clique com o botão direito no **Step Dummy** e selecione **Preview**. Veja as linhas que deram erro.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Agora é com você. Veja a figura abaixo e termine o trabalho prático.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Observe o conteúdo de cada Step.

**Step “Select values 2”**

A saída desse Step é:

Interface gráfica do usuário, Tabela

Descrição gerada automaticamente com confiança média

**Step Sort rows**

Interface gráfica do usuário, Aplicativo, Tabela

Descrição gerada automaticamente

**Step Unique rows**



Esta transformação tem como função, remover linhas duplicadas do fluxo de entrada e filtrar apenas as linhas exclusivas para seguimento no fluxo.

Pré-Requisito: para entregar um resultado correto, o fluxo de entrada **deve ser classificado em uma etapa anterior**, caso contrário, apenas as linhas duplas consecutivas serão analisadas e filtradas. Para isso, vamos utilizar o **Step Sort rows**.

Interface gráfica do usuário, Aplicativo, Tabela

Descrição gerada automaticamente

**Step Text file output 2 – Resultado 2**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

**DIM\_Producao (Step Table output)**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

O **Step Table output** carrega dados em uma tabela do banco de dados. Esse Step é equivalente ao operador SQL INSERT e é uma solução quando você só precisa inserir registros. Se você quiser apenas atualizar linhas, use Update. Para executar os comandos INSERT e UPDATE, consulte o **Step Insert / Update**.

Esta etapa fornece opções de configuração para uma tabela de destino e opções relacionadas ao desempenho, como Commit Size e Use **batch update** para inserções. Existem configurações de desempenho específicas para um tipo de banco de dados que podem ser definidas nas propriedades JDBC da conexão com o banco de dados.

Use um gerenciador de Banco de Dados. No nosso caso poderia ser o MySql Workbench ou MySql Xampp, porém você pode usar outro banco de dados se conseguir estabelecer a conexão.

**Connection**: conexão com o banco de dados.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

**Target schema**: nome do schema.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

**Target table**: nome da tabela.

Para esse caso informe o nome da tabela e clique no botão Navegar. Se a conexão estiver ok, ele vai mostrar o caminho no banco de dados.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Se a tabela não existir ainda, clique no **botão SQL**. Irá abrir uma nova janela com o script de criação da tabela. Clicar no **botão** **Execute** e a tabela será criada, mas não o faça ainda. Cabe uma observação aqui.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

O PDI se baseia nos primeiros registros do fluxo de dados para sugerir o script de criação da tabela no banco de dados. Portanto ele vai sugerir que o campo **NM\_MUNICIPIO\_PROGRAMA\_IES** seja criado com **tamanho 21**. O problema é que mais ao final dos dados, nas linhas finais, tem registros onde esse campo **NM\_MUNICIPIO\_PROGRAMA\_IES** tem um valor bem maior, **próximo de 50 caracteres**. Assim você precisa mudar esse tamanho para não receber erros durante a transformação.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Rode a transformação.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Clique no **Step DIM\_Producao** com o botão direito e selecione **Preview** para ver os dados.

Interface gráfica do usuário, Aplicativo, Tabela, Excel

Descrição gerada automaticamente

Rode o comando SQL abaixo no Banco de Dados para verificar os dados que foram inseridos na tabela.

*select \* from tp.`dim-programa`;*

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Agora vamos carregar uma dimensão UF a partir dos dados do arquivo de Resultado 2.

* Select values 3
* Table output (UF)

Uma imagem contendo Diagrama

Descrição gerada automaticamente

Ao fazer um **Preview** na **DIM\_UF**, você verá muita repetição de UFs.

Rode os seguintes scripts no MySql:

* Vamos limpar a tabela:

*truncate table tp.uf;*

* Vamos criar uma primary key para não termos repetição de uf:

*alter table tp.uf*

*add primary key (sg\_uf\_programa);*

Rode a transformação e você verá o erro de integridade referencial.

Diagrama

Descrição gerada automaticamente com confiança média

Texto

Descrição gerada automaticamente

Insira um Step **Dummy** e veja o que acontece (Preview).

Diagrama

Descrição gerada automaticamente

Agora vamos fazer a **orquestração** da transformação e para isso vamos criar um arquivo de **Job** conforme imagem abaixo.

Interface gráfica do usuário, Aplicativo, Word

Descrição gerada automaticamente

Reproduza o Job, conforme Steps da imagem abaixo.

Linha do tempo

Descrição gerada automaticamente com confiança média

No Step **Transformation**, linkar a transformação trabalhada.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

**Rodar a transformação pelo Job.**

**2) Prática Apache HOP**

Para iniciar, vamos converter o trabalho prático executado no Pentaho PDI (Spoon.bat) para o Apache Hop.

Crie um novo projeto.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Preencha e nome e o diretório onde será armazenado o projeto.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Clique no ícone Hop.

Interface gráfica do usuário, Aplicativo, Word

Descrição gerada automaticamente

Selecione “Import from Kettle/PDI”.

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Informe o diretório onde estão os arquivos ktr (transformação) e kjb (job) do Pentaho PDI.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Informe o projeto e faça a importação.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Clique no ícone abrir arquivo

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Informe o diretório onde foi feita a importação e abra os dois arquivos – hpl (transformação) e hwf (job).

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

A transformação vai dar erro de conexão por conta do driver MySQL.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Precisa baixar o conector do MySQL para o Apache Hop. É possível fazer isto no site <https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-j-8.1.0.zip>

Abra o arquivo ***mysql-connector-j-8.1.0.zip***.

Selecione o arquivo ***mysql-connector-j-8.1.0.jar*** e copie para dentro do Apache Hop, ...\plugins\databases\mysql\lib.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Feche o Apache Hop e abra novamente testando a conexão com o MySQL. Se possível, crie uma conexão nova.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

**Agora já é possível rodar o Job no Apache Hop.**

**Linha do tempo

Descrição gerada automaticamente**