Exercício 1 – Criando a transformação “Hello World”

1 – Crie uma pasta chamada `pdi_labs` no seu computador.

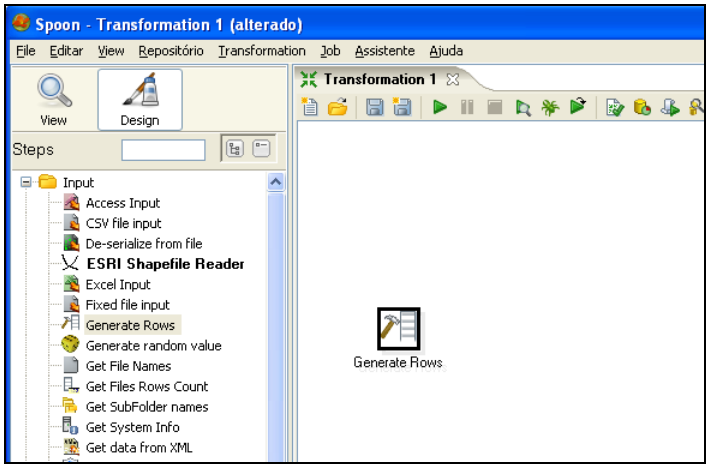
2 – Abra o Spoon.

3 – A partir do menu principal escolha **Arquivo -> Nova Transformação**. Será mostrada a transformação *Transformation 1* na área de trabalho.

Obs.: Caso deseje, utilize as teclas de atalho CTRL-N

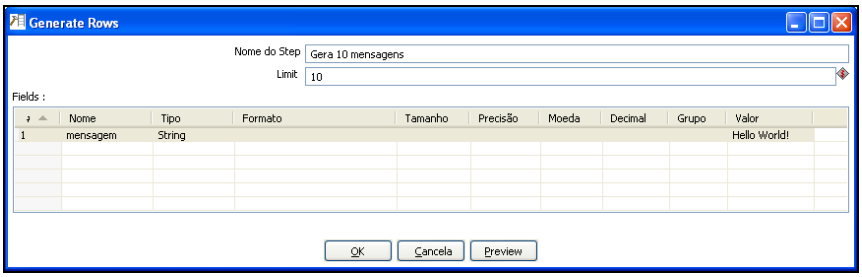
4 – Do lado direito da tela é mostrada a árvore de *steps*. Clique na opção *Design* e Expanda a opção *Input*.

5 – Arraste e solte o ícone do step *Generate Rows* para a área de trabalho da transformação.



6 – Dê um duplo-clique para abrir o step e digite os dados abaixo e clique OK em seguida:

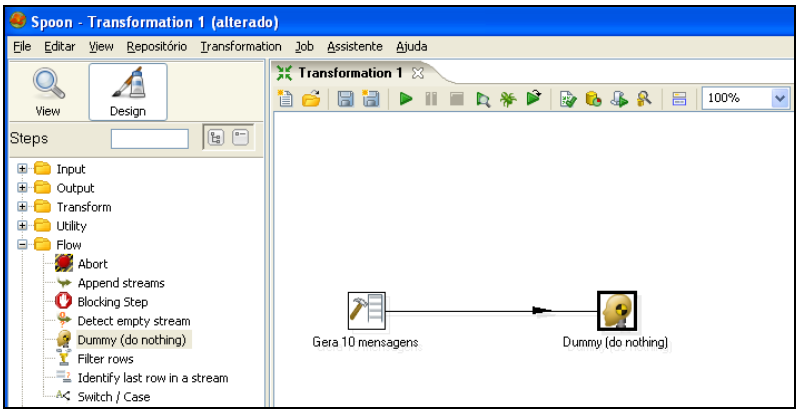
- Nome do step: Gera 10 mensagens
- Em Fields:
  - Nome: mensagem
  - Tipo: String
  - Valor: Hello World!



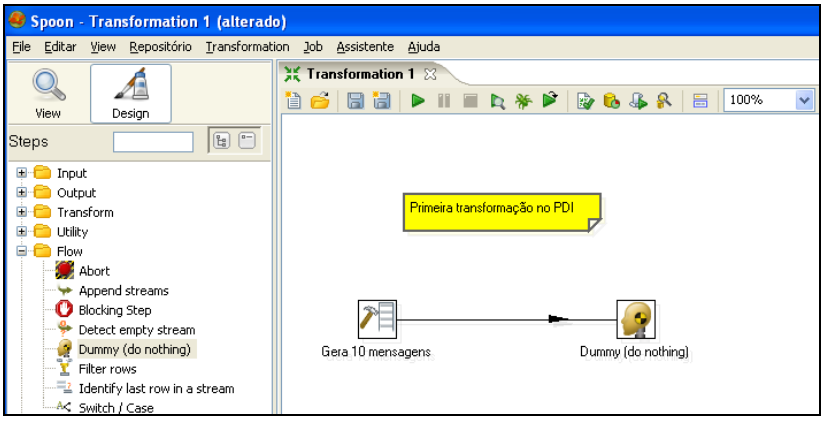
Obs.: Você também pode abrir o step para edição clicando com o botão direito e escolhendo a opção *Editar Step*.

7 – Na árvore de steps, expanda a opção *Flow* e arraste e solte o ícone do step *Dummy* para a área de trabalho da transformação.

8 – Vamos criar um *hop* para ligar os dois steps. Clique com o botão direito no step *Generate Rows*, segure a tecla *Shift* e arraste o cursor do mouse para o step *Dummy*.

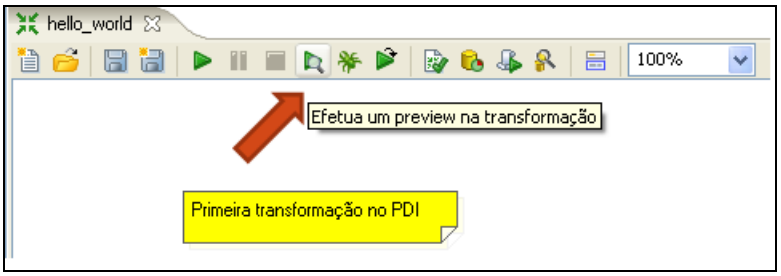


9 – Vamos adicionar uma nota à transformação. Clique com o botão direito em qualquer parte da área de trabalho da transformação e escolha a opção *Nova Nota*. Digite o texto da nota.

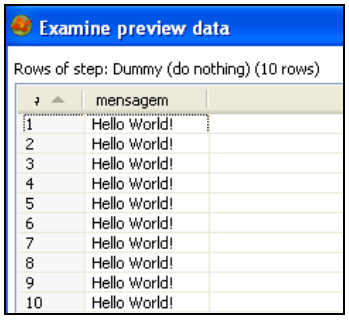


10 – Salve a transformação na pasta *pdi_labs* com o nome *hello_world*. Será gerado um arquivo XML com extensão .ktr.

11 – Podemos rodar uma prévia da transformação, antes de sua execução real. Clique com o botão direito no step *Dummy* e escolha a opção *Preview*.
Obs.: Você poderá rodar a previsão da transformação clicando no step desejado e, em seguida, clicando no ícone *Preview* da barra de opções.



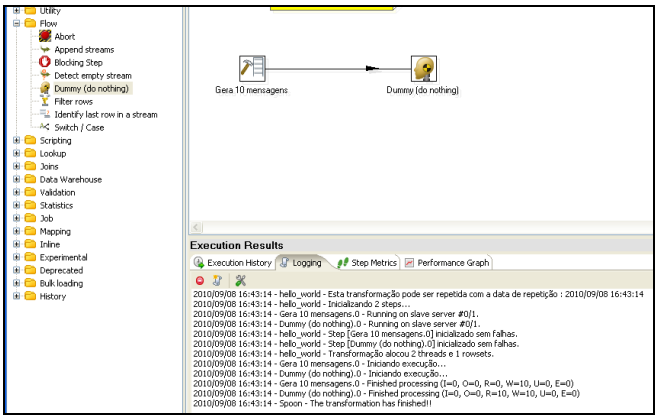
12 – Após carregar a janela de diálogo de *Preview*, clique no botão *Quick Launch* para visualizar o resultado da transformação.



13 – O que foi feito nessa transformação? O step *Generate Rows* simplesmente gera várias linhas de registros, de acordo com os parâmetros informados. Experimente editar o step para acrescentar outros campos e mudar o tamanho do conjunto de registros gerados. O step *Dummy* apenas recebe os registros gerados e não faz nada.

14 – Um pequeno detalhe: nós não executamos a transformação, apenas visualizamos uma prévia de sua execução. Para executar a transformação clique no botão *Run* e após abrir a janela de diálogo clique no botão *Launch*.

15 – O resultado real da execução aparece na aba do log. As métricas de execução aparecem na aba da janela do log. O nível de detalhe do log de execução pode ser configurado na tela anterior.
Obs.: na execução de grandes volumes de dados recomenda-se deixar o nível do log como *Basic*.



Exercício 2 – Criando a transformação "Hello World Fulano"

1 – Vamos expandir a transformação do exercício anterior para obter dados de uma fonte, fazer uma pequena transformação nos dados e armazenar o resultado em um arquivo texto. Crie um arquivo texto com os dados abaixo:

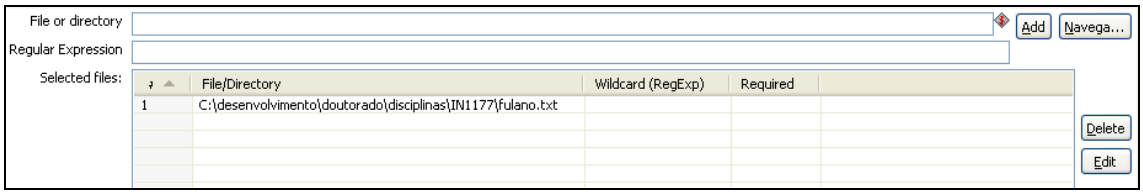
```
Fulano
Sicrano
Beltrano
```

Salve o arquivo na pasta *pdi_labs* com o nome *fulano.txt*.

2 – Crie uma nova transformação e salve-a com o nome de *hello_world_fulano.ktr*. O objetivo dessa transformação é ler o arquivo texto acima, adicionar a frase "Hello World" e uma sequência para cada frase criada. Ao final, todas as frases criadas serão gravadas em um arquivo texto.

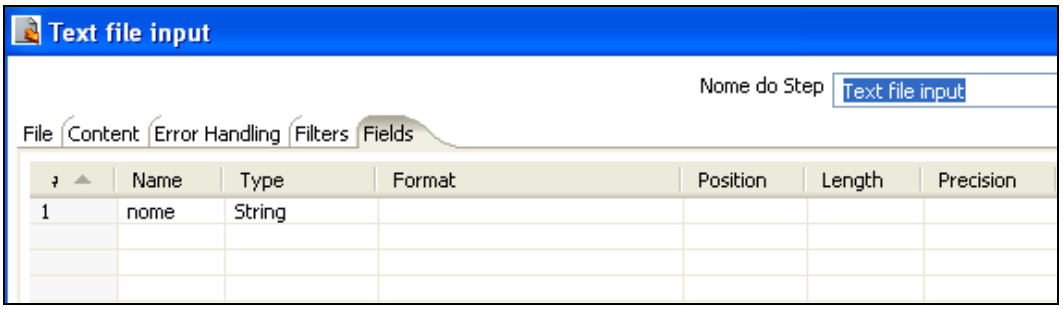
3 – Adicione um step *Text file input* da opção *Input*. Edite o step com as seguintes informações:

- na aba *File*, campo *File or directory*, clique no no botão *Navegar* para localizar o arquivo *fulano.txt*. Em seguida, clique no botão *Add* para colocar o arquivo na lista.



- Na aba *Content*, desmarque a opção *Header*. Se você não fizer isso, o Spoon entenderá que a primeira linha do arquivo texto é um cabeçalho com o nome de um ou mais campos.

- Na aba Fields, digite o nome do campo (*nome*) e o seu tipo (*String*). Clique OK.

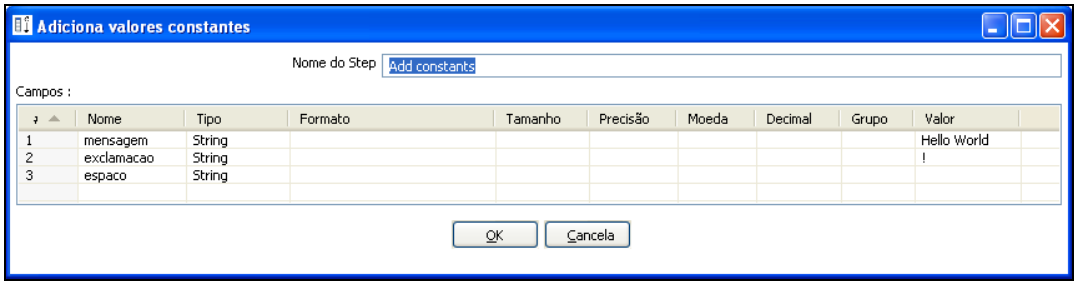


4 – Adicione um step *Add constants*, do tipo *Transform*. Edite o step com as seguintes informações nos campos *Nome*, *Tipo* e *valor*:

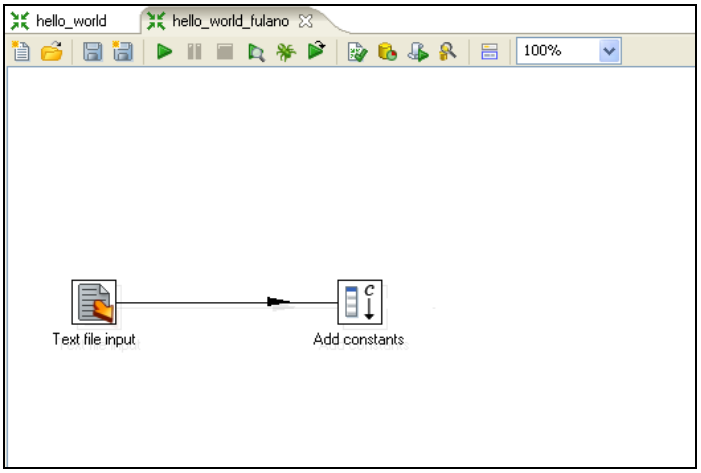
- Nome: mensagem; Tipo: String; Valor=Hello World

- Nome: exclamação; Tipo: String; Valor=!

-Nome: espaço; Tipo: String; Valor=” “ (espaço em branco)



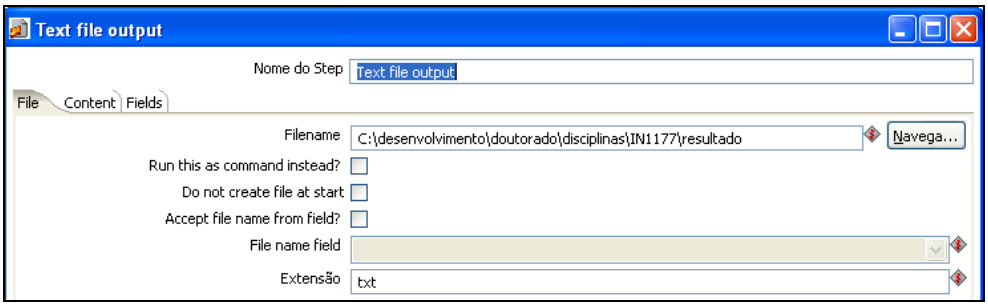
5 – Crie um hop ligando os dois steps.



6 – Ainda nos steps do tipo *Transform*, adicione um *Add Sequence*. Em seguida, crie um hop ligando o *Add constants* ao *Add sequence*. Edite esse step para ver os seus parâmetros (não vamos alterá-los para esse exercício).

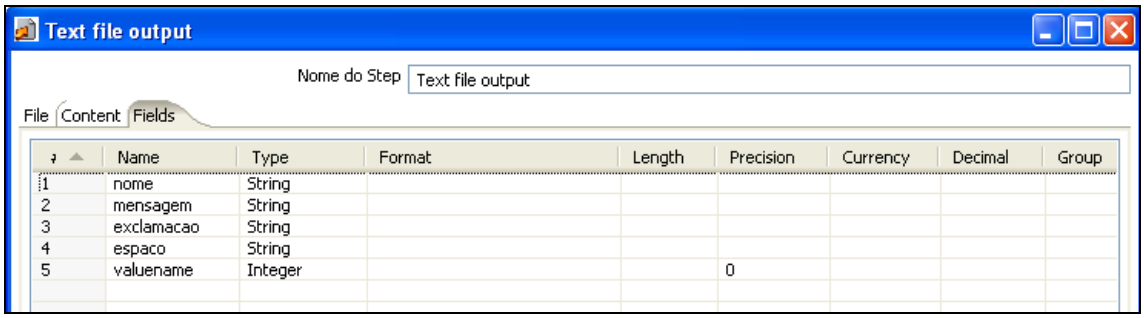
7 – Entre nas opções de steps do tipo *Output* e adicione um step *Text file output*. Em seguida, crie um hop ligando o step *Add sequence* a *Text file output*. Edite as configurações do step:

- Na aba *File*, digite o caminho de um arquivo de texto com o nome *resultado* (o step adiciona por default a extensão .txt) no campo *Filename*.

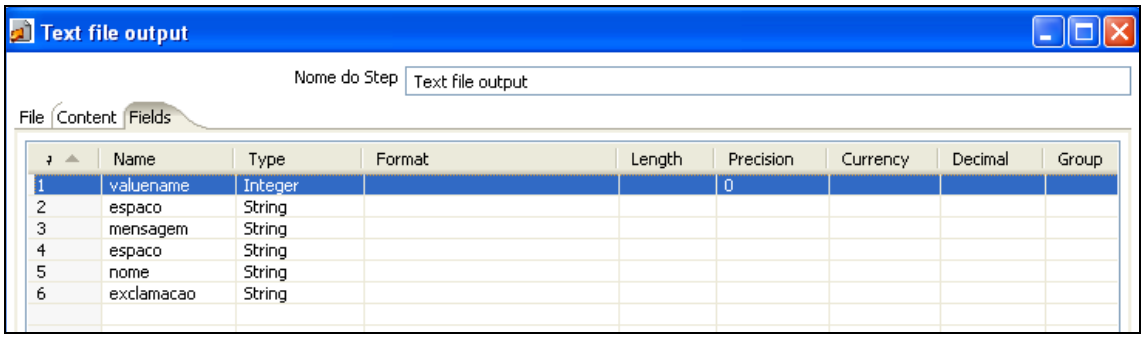


- Na aba *Content*, limpe o conteúdo do campo *Separator* (caso contrário, o step colocará um separador na saída do arquivo) e desmarque a opção *Header*.

- Na aba Fields, clique no botão *Obtém campos*. Se tudo foi feito corretamente até agora, serão mostrados os campos de acordo com a figura abaixo. Observe que os campos foram gerados pelos steps anteriores.



Queremos que seja gravada no arquivo a mensagem "<valuename> <mensagem><espaço> <nome><exclamacao>". Para que isso ocorra devemos modificar a ordem dos campos. Clique com o botão direito em cima do campo *valuename* e escolha a opção *move up*. Repita a operação até que o campo *valuename* seja o primeiro da lista. Repita a operação com os demais campos. A figura abaixo mostra a disposição final dos campos.



8 – Salve a transformação e execute em seguida. O resultado gerado é um arquivo texto com as seguintes linhas:

```
1 Hello World Fulano!
2 Hello World Sicrano!
3 Hello World Beltrano!
```

Exercício 3 – Criando uma conexão com um banco de dados.

1- Nos exercícios anteriores mostramos como obter dados a partir de uma fonte de dados baseada em arquivos texto. Para os exercícios futuros precisaremos extrair, transformar e carregar dados em tabelas de um banco. Para facilitar o processo de aprendizagem, vamos utilizar o SGBD open source Apache Derby. Faça o download da última versão através da url http://db.apache.org/derby/derby_downloads.html.

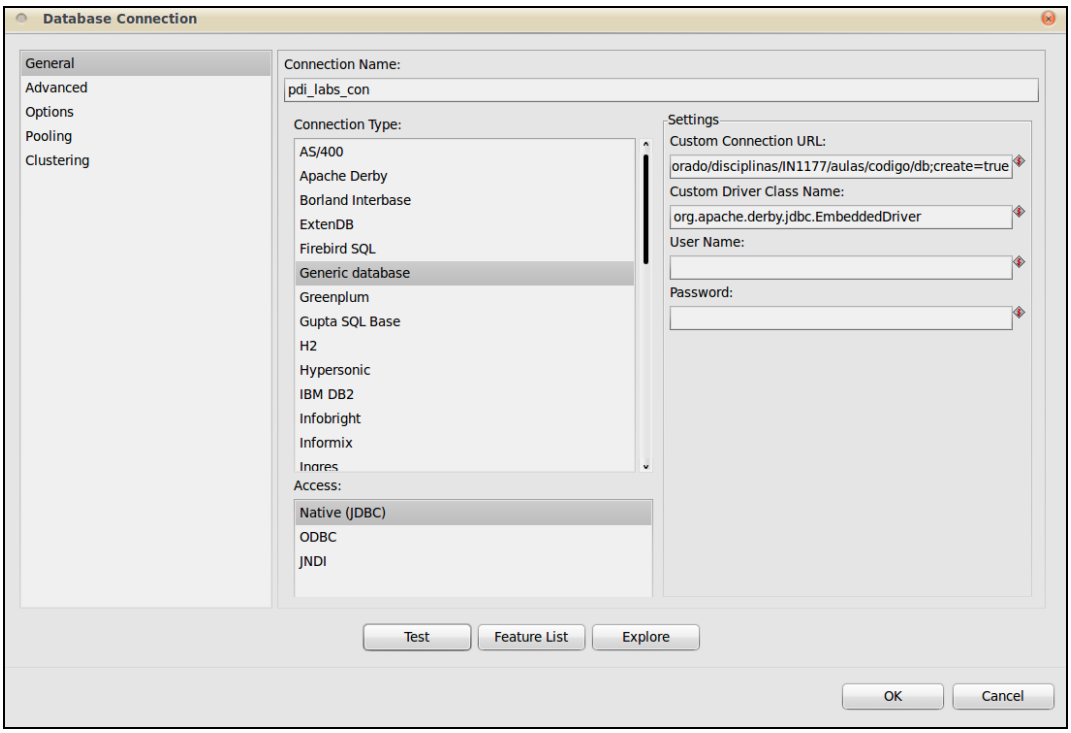
2- Descompacte o arquivo em uma pasta no seu computador. Em seguida, copie o arquivo */lib/derby.jar* para a pasta */JDBC/libext* de sua instalação do PDI.

3- Abra o Spoon e carregue a transformação do exercício anterior. Em seguida, clique no botão *View* na barra lateral. Expanda a aba *Conexões*.

4- Com o botão direito, clique em cima da opção *Conexões* e escolha *Novo*.

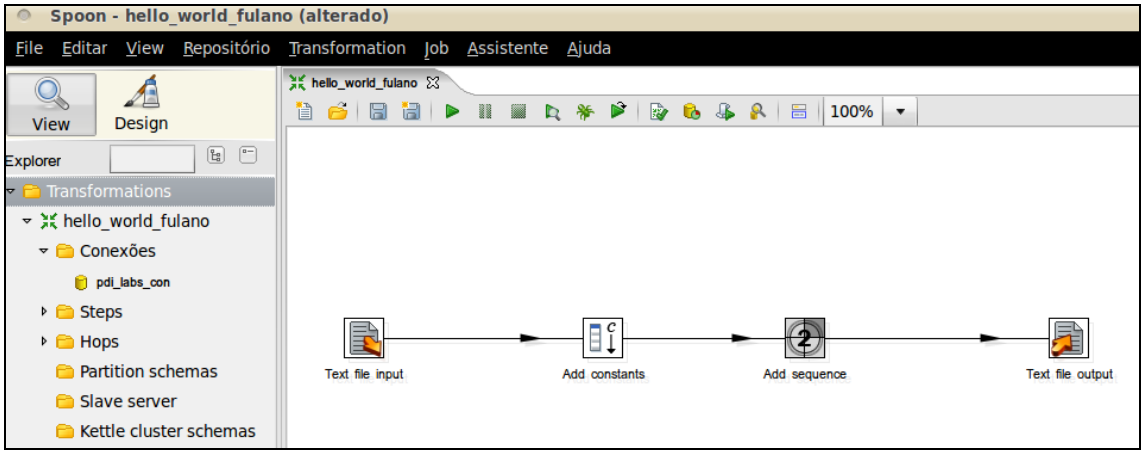
5- Na aba General, digite os seguintes parâmetros:

- Connection type: Generic database
- Access: Native (JDBC)
- Custom Connection Url: jdbc:derby:<path_do_banco>/<banco>;create=true
- Custom Driver Class Name: org.apache.derby.jdbc.EmbeddedDriver



6– Clique no botão Test para verificar se a conexão com o banco está OK. Em seguida, retire o texto "`;create=true`" da Url de conexão (caso contrário, as tabelas serão apagadas e criadas a cada inicialização da conexão). Clique em OK.

6- Podemos tornar a conexão visível para todas as transformações e Jobs de nossa instalação do PDI. Para isso, clique com o botão direito em cima da conexão criada e escolha a opção *Share*. Note que o texto da conexão está em negrito agora.



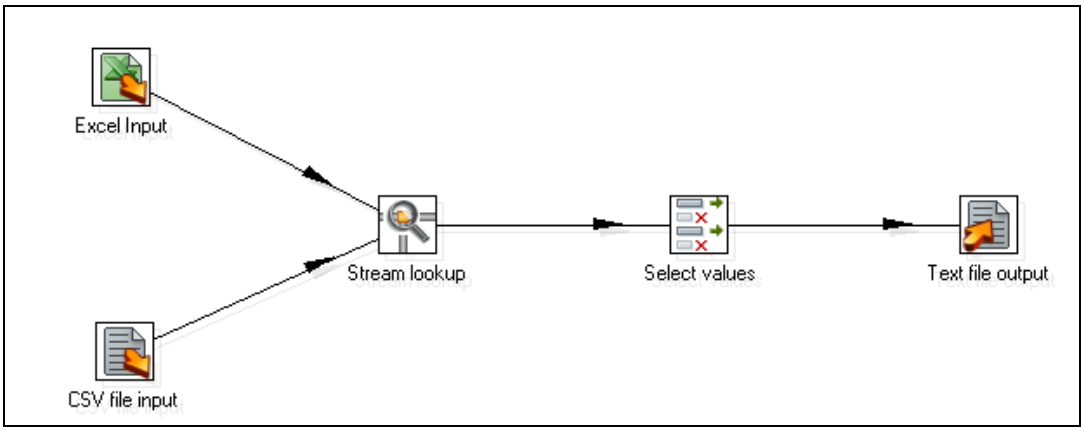
7- O banco criado está vazio, sem nenhuma tabela. Mais adiante vamos utilizar essa e outras conexões para a modelagem dimensional e a execução do processo de ETL.

Exercício 4 – Extraindo dados de um arquivo texto, realizando uma transformação e carregando o resultado em um arquivo texto.

1 – Para realizar esse exercício você precisará de dois arquivos armazenados na pasta *Bases*: *capes.xls* e *ies.csv*. O primeiro contém informações sobre a avaliação dos programas de Pós-Gradução das Instituições de Ensino Superior (IES) do país. O segundo arquivo é uma tabela de lookup contendo o código e a sigla das IES's. Vamos mostrar a extração de dados a partir de dois arquivos (Excel e CSV), uma pequena transformação e o carregamento dos dados em um arquivo texto. Abra a planilha e veja que o campo *ies* possui a sigla da instituição. Queremos gravar em um arquivo texto parte dos dados da planilha e o código da instituição no lugar de sua sigla.

Para iniciar, abra o Spoon e crie uma nova transformação.

2 – Abra a categoria *Input* e adicione os steps *Excel Input* e *CSV file input*. Em seguida, expanda a categoria *Lookup* e adicione o step *Stream lookup*. Da categoria *Transform*, adicione o step *Select values*. Da categoria *Output*, adicione o step *Text file output*. Por fim, crie os hop's para conectar os steps, de acordo com a figura abaixo.



3 – Edite o step *Excel input* com os seguintes parâmetros:

- Aba *Files*
    - *File or directory*: localize o arquivo *capes.xls* com o botão *Navegar*. Em seguida, clique em *Add* para adicionar o arquivo ao grid. Para ter certeza que o arquivo foi localizado, clique no botão *Show filename(s)*.
- Aba *Sheets*
    - Clique no botão *Get sheetname* e escolha a planilha desejada. Se o nome da planilha não aparecer na lista, reveja os parâmetros da aba *Files*.
- Aba *Content*
    - Certifique-se que o campo *Header* esteja marcado (vamos precisar dele na próxima aba).
    - Esse arquivo foi gravado no Linux! Mude o campo *Encoding* para UTF-8.
- Aba *Fields*
    - Clique no botão *Get fields from header now* e veja todos os campos disponíveis no arquivo.
    - Dê uma olhada nos dados que serão extraídos do arquivo, clicando no botão *Preview rows*. Clique *Ok* e salve a transformação.

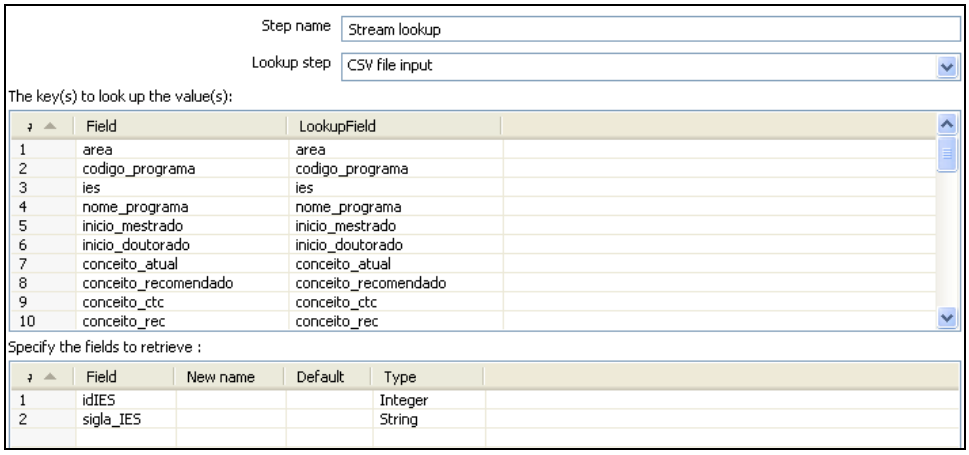
4 – Edite o step *CSV file input* com os seguintes parâmetros:

- *Filename*: localize o arquivo *ies.csv* com o botão *Navegar.*
- *Delimiter*: ; (ponto-e-vírgula)
- Desmarque a opção *Lazy conversion*.
- Clique no botão *Obtém campos* e veja os campos que serão lidos.
- Clique no botão *Preview* para visualizar uma amostra dos dados.
    - Na grade com os campos, diminua o tamanho do campo *idIES* para 1 (propriedade *Length*). Rode novamente o *preview*.
    - Retire o símbolo da moeda (R$) da propriedade *Currency*.

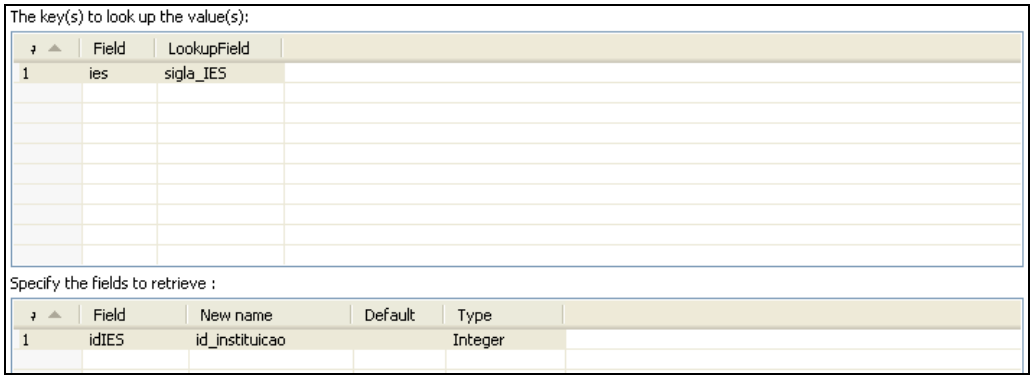
5 – Antes de editar o step *Stream lookup*, dê uma olhada no fluxo de registros de entrada. Clique com o botão direito em cima do step e escolha a opção *Mostra campos de entrada*. Deverão ser exibidos 41 campos (39 da planilha e 2 do arquivo texto).

6 – Edite o step *Stream lookup* com os seguintes parâmetros:

- Lookup step: escolha o step *CSV file input*.
- Clique nos botões Get fields e Get lookup fields. As grades de campos deverão ter a configuração da figura abaixo.



- Na grade de cima, remova todos os campos, deixando apenas o campo *ies*. Em *Lookup Field* escolha o campo *sigla_IES*.
- Na grade de baixo, remova o campo *sigla_IES*. No campo *idIES*, digite o valor *id_instituicao* na propriedade *New name*. As grades deverão ter a configuração abaixo. Salve a transformação.



- Na área de trabalho, clique com o botão direito em cima do step *Stream lookup* e escolha a opção *Mostra campos de saída*. Observe que temos agora 40 campos (o último é o campo que denota o *id* da instituição).

7 – Suponha que não precisamos de todos os campos vindos da planilha. Além disso, queremos modificar o nome dos campos que serão carregados ao final do processo de transformação. Edite o step *Select values* com os seguintes parâmetros:

- Aba *Meta-data*
    - *Fieldname*: escolha o campo *area* digite *id_area* na propriedade *Rename to* e escolha *Integer* na propriedade *Type*.
    - Repita a operação com o campo *codigo_programa*, com o nome *id_programa*. A grade deverá ter a configuração abaixo.

Fields to alter the meta-data for :

#	Fieldname	Rename to	Type
1	area	id_area	Integer
2	codigo_programa	id_programa	-

- Aba *Remove*
    - Clique no botão *Get fields to remove* e **exclua** os seguintes campos:
        - area, codigo_programa, ies, nome_programa, inicio_mestrado, inicio_doutorado, conceito_recomendado e id_instituicao.
        - A lista de campos que queremos excluir deverá ter a configuração abaixo.

Fields to remove :

#	Fieldname
1	conceito_atual
2	conceito_ctc
3	conceito_rec
4	q2_q5
5	itens_q2_q5
6	itens_q3_q4
7	docentes_ano_2001_2003
8	docentes_ano_2004_2006
9	docentes_ano_2001_2006
10	Teses_2001_2003

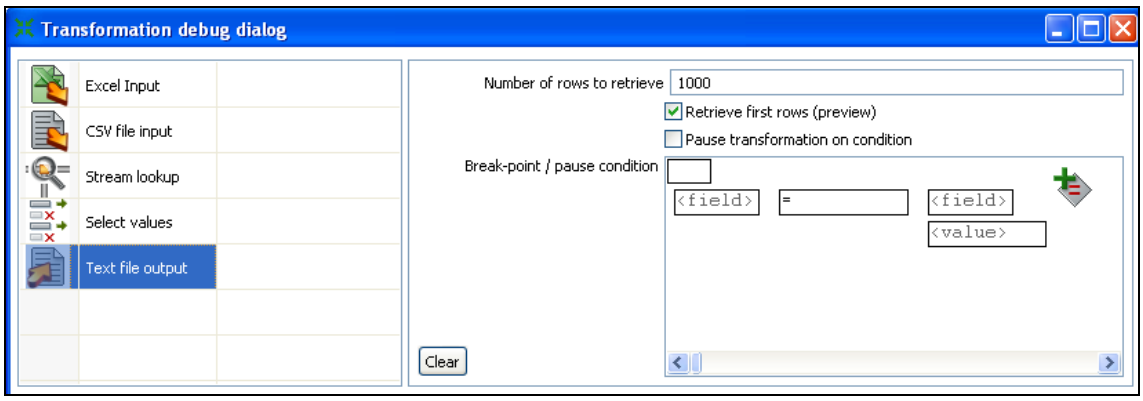
Salve a transformação e veja os campos de saída do step, clicando com o botão direito em cima dele.

8 – Vamos editar o step *Text file output* com os seguintes parâmetros:

- Aba *File*
    - *Filename*: <path>\resultado
- Aba *Fields*
    - Clique no botão *Obtém campos* e veja os campos que serão gravados. Alterações de formato, tamanho, etc. poderão ser feitas diretamente na grade.
    - Clique no botão *Minimal width* e veja que o step fornece um formato padrão para os campos.

9 - Salve a transformação. Na área de trabalho, clique com o botão direito em cima do step *Text file output* e escolha a opção *Preview*. Na janela de diálogo, marque a opção

Text file output, conforme a figura abaixo. Se tudo foi configurado corretamente, será mostrada uma amostra dos dados que serão gravados.



10 – Execute a transformação e veja o arquivo gerado. O que fizemos nessa transformação?

- Mostramos como extrair dados de dois arquivos com formatos diferentes (.xls e .csv).
- Fizemos duas transformações nos dados extraídos: trocamos o nome da IES por seu código e retiramos os campos que não queríamos gravar. Além disso, alteramos os metadados de dois campos.
- Carregamos o resultado da transformação em um arquivo texto.

Exercício 5 – Extraindo dados de um conjunto de arquivos

1 – Para esse exercício vamos utilizar as planilhas armazenadas na pasta */pdi_labs/base/planilhas*. Crie uma nova transformação e copie/cole os steps e hops da transformação do exercício anterior.

2 – Edite o step *Excel input*. Na grade de arquivos selecionados da aba *Files*, marque o arquivo *capes.xls* e clique no botão *Delete*.

3 – Acesse a pasta *planilhas* e adicione todos os arquivos na grade usando o botão *Add* (Ok...não precisa adicionar todos. São 45 arquivos ☺). Verifique se o step carregou os arquivos clicando no botão *Show filename(s)*.

Selected files:

#	File/Directory	Wildcard (RegExp)	Required
1	E:\pdi_labs\bases\planilhas\01_Mat_Estat.xls		
2	E:\pdi_labs\bases\planilhas\02_Computacao.xls		
3	E:\pdi_labs\bases\planilhas\03_Fisica_Astronomia...		
4	E:\pdi_labs\bases\planilhas\04_Quimica.xls		
5	E:\pdi_labs\bases\planilhas\05_Geociencias.xls		
6	E:\pdi_labs\bases\planilhas\06_Bio1.xls		
7	E:\pdi_labs\bases\planilhas\07_Ecologia.xls		
8	E:\pdi_labs\bases\planilhas\08_Bio2.xls		
9	E:\pdi_labs\bases\planilhas\09_Bio3.xls		
10	E:\pdi_labs\bases\planilhas\10_Eng1.xls		
11	E:\pdi_labs\bases\planilhas\11_Artes.xls		
12	E:\pdi_labs\bases\planilhas\12_Eng2.xls		
13	E:\pdi_labs\bases\planilhas\13_Eng3.xls		
14	E:\pdi_labs\bases\planilhas\14_Eng4.xls		
15	E:\pdi_labs\bases\planilhas\15_Med1.xls		
16	E:\pdi_labs\bases\planilhas\16_Med2.xls		
17	E:\pdi_labs\bases\planilhas\17_Med3.xls		
18	E:\pdi_labs\bases\planilhas\18_Odontologia.xls		
19	E:\pdi_labs\bases\planilhas\19_Farmacia.xls		

4 – Na aba *Sheets* exclua a planilha do exercício anterior e clique em *Get sheetname(s)* para pesquisar pelas planilhas de **todos** os arquivos selecionados no passo anterior. Escolha a planilha *plan1*.

5 - Na aba *Content* certifique-se que o campo *Header* esteja marcado e que o *Encoding* do arquivo seja UTF-8 ou ISO-8859-1 (experimente colocar outra codificação e veja o que irá ocorrer).
6 – Na aba *Fields* clique no botão *Get fields from header now*, lembrando-se de limpar a lista de campos do exercício anterior (confirme na janela de diálogo). Salve a transformação e execute-a. Veja o resultado do arquivo gerado pela transformação.

Exercício 6 – Extraindo dados de um conjunto de arquivos, usando expressões regulares.

1 – A solução adotada no exercício anterior pode ser refeita para evitar a inclusão de cada arquivo manualmente (imagine uma aplicação real com milhares de arquivos). Abra a transformação do exercício anterior e edite o step *Excel input*.

2 – Na aba *Files*, exclua todos os arquivos da grade *Selected files*. No *campo File ou directory* digite o caminho para a pasta onde as planilhas estão armazenadas (ex.: E:\pdi_labs\bases\planilhas\). Adicione o caminho à lista de arquivos clicando no botão *Add*.

3 – No campo *Selected files*, digite a seguinte expressão na propriedade *Wildcard*:
`.*\.xls`

Selected files:	#	File/Directory	Wildcard (RegExp)	Required
	1	E:\pdi_labs\bases\planilhas\	.*\.xls	

4 – Para garantir que os arquivos serão lidos usando a expressão regular, clique no botão *Show filename(s)* e veja a lista de arquivos que serão lidos. Salve a transformação e execute-a, observando o arquivo gerado.

Exercício 7 – Selecionando, alterando campos e gerando a saída em uma planilha.

1 – Vamos melhorar o cabeçalho do arquivo gerado. Abra o exercício anterior e edite o step *Select values*. Remova todos os campos selecionados nas abas *Remove* e *Meta-data*.

2 – Na aba *Select & Alter* digite os nomes dos campos, de acordo com a figura abaixo. Salve a transformação e veja os campos de saída do step.

Select & Alter | Remove | Meta-data

Fields :

#	Fieldname	Rename to	Length	Precision
1	codigo_programa	Código do Programa		
2	nome_programa	Nome do Programa		
3	area	Código da Área		
4	id_instituicao	Código da Instituição		
5	ies	Instituição		
6	inicio_mestrado	Início Mestrado		
7	inicio_doutorado	Início Doutorado		
8	conceito_recomendado	Conceito CAPES		

3 – Clique no step File text output e apague-o. Da categoria *Output*, crie um step do tipo *Excel output*. Crie um hop ligando os steps *Select values* e *Excel output*.

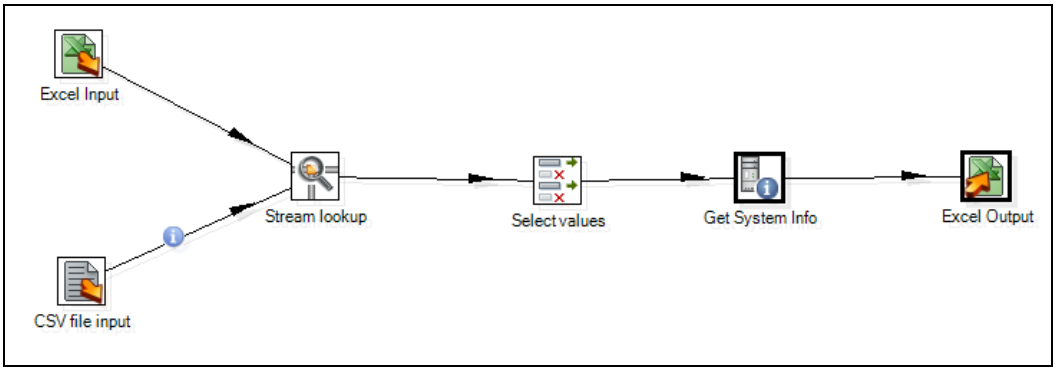
4 – Edite o step *Excel output* com os seguintes parâmetros:

- Aba *File*
  - *Filename*: o caminho e nome da planilha gerada.
- Aba *Fields*
  - Clique nos botões *Obtém campos* e *Minimal Width*
- (opcional) Aba *Content*
  - Explore as opções da aba (Split, sheet name, protect, templates)

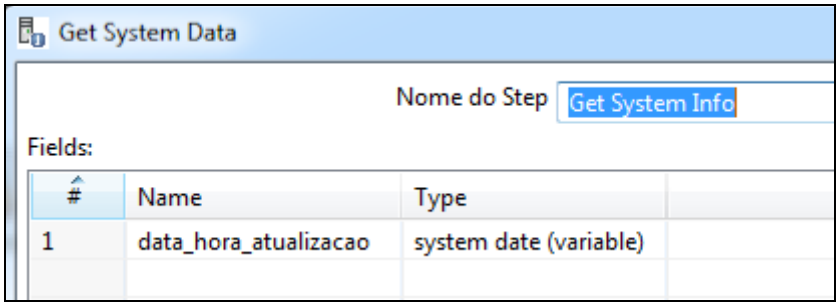
5 – Salve a transformação e veja a sua execução.

Exercício 8 – Obtendo informações do sistema.

1 – Vamos acrescentar uma informação correspondente à data e hora em que o registro foi gravado no arquivo. Abra a transformação do exercício anterior e crie um step *Get system info* da categoria *Input*. Coloque esse step entre o *Select values* e o *Excel output*. Crie os hops ligando os steps, de acordo com a figura abaixo.

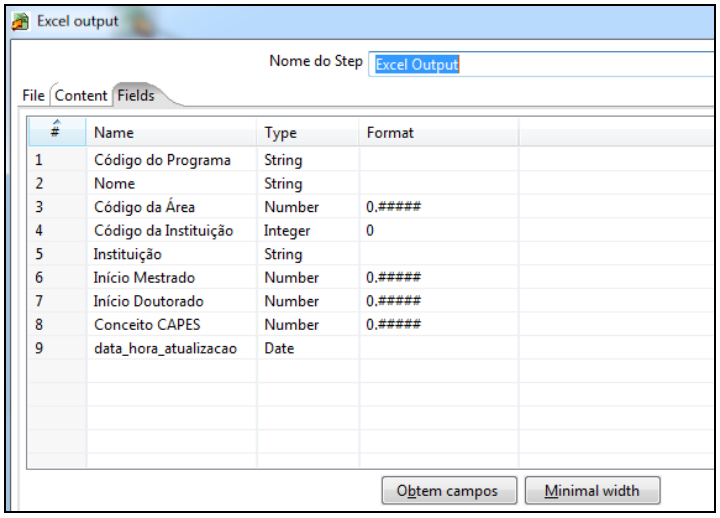


2 – Edite o step *Get system info* com os seguintes parâmetros:



3 – Edite o step *Excel output*, com as seguintes alterações:

- Na aba Fields, clique no botão Obtém campos, e verifique se o campo data_hora_atualizacao será exibido.



4 – Salve a transformação e execute-a. Veja o arquivo gerado.

Exercício 9 – Aplicando formatos para datas e números.

1 – Você deve ter notado que os registros gravados nos arquivos dos exercícios não possuem uma formatação adequada. Antes de colocar os formatos corretos devemos escolher qual step deverá formatar os valores. Por uma questão de coesão, os steps mais adequados são aqueles relacionados às atividades de transformação. Vamos editar o step *Select values* para aplicar os formatos.

2 – No step *Select values*, marque todas as linhas da aba *Select & alter*, recorte e cole na aba *Meta-data*. Complete as propriedades com os valores abaixo.

Select / Rename values

Step name: Select values

Select & Alter | Remove | Meta-data

Fields to alter the meta-data for :

#	Fieldname	Rename to	Type	Length	Precision	Binary to Normal?	Format
1	codigo_programa	Código do Programa	-			N	
2	nome_programa	Nome	-			N	
3	area	Código da Área	Integer	1		N	#
4	id_instituicao	Código da Instituição	Integer	1		N	#
5	ies	Instituição	-			N	
6	inicio_mestrado	Início Mestrado	Integer	4		N	
7	inicio_doutorado	Início Doutorado	Integer	4		N	
8	conceito_recomendado	Conceito CAPES	Integer	1		N	#

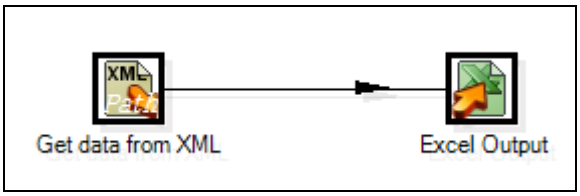
3 – Edite o step *Excel output*, abrindo a aba *Fields*. No campo *data_hora_atualizacao* digite o seguinte formato: `dd/MM/yyyy HH:mm:ss`

Se existir, limpe o formato dos demais campos.

4 – Salve a transformação, execute-a e observe o arquivo gerado.

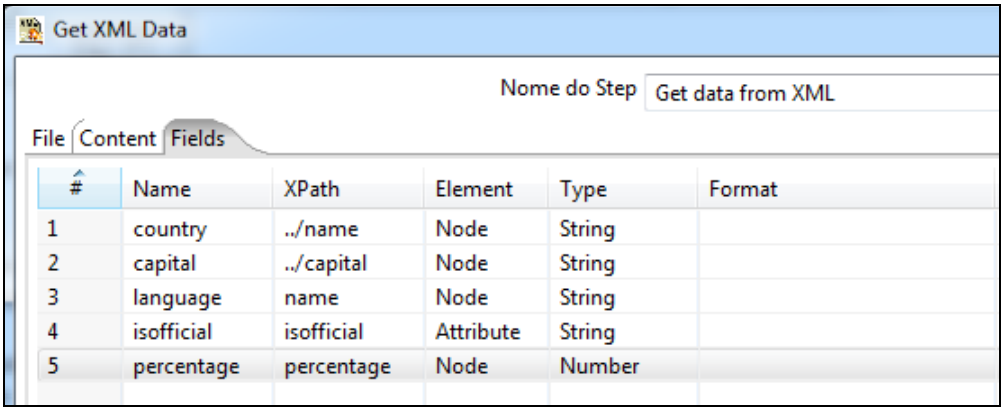
Exercício 10 – Extraindo uma lista com dados de países de um arquivo XML.

1- Para esse exercício vamos utilizar o arquivo *countries.xml* que encontra-se na pasta */pdi_labs/bases/xml*. Abra o arquivo e observe a sua estrutura.
2- Crie uma nova transformação e adicione os steps *Get data from XML* e *Excel output*. Crie um hop ligando os steps.



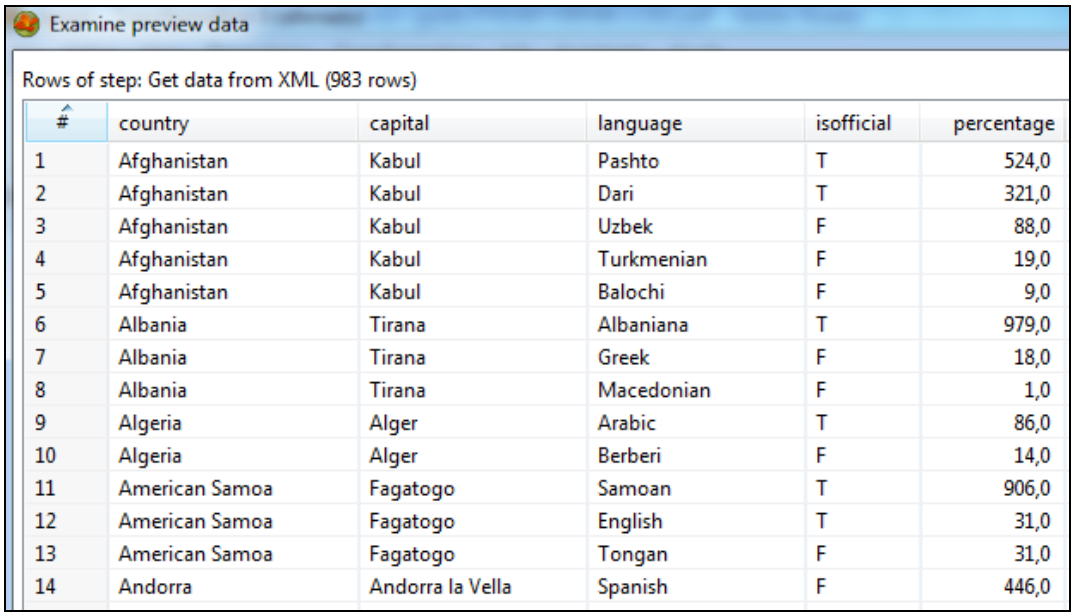
3- Edite o step Get data from XML, com os seguintes parâmetros:

- Aba File
  - *File or directory*: encontre o arquivo *countries.xml* e adicione à lista.
- Aba *Content*
  - Clique no botão *Get Xpath nodes* e selecione */world/country/language*
- Aba Fields
  - Preencha a grade de acordo com a figura abaixo.



#	Name	XPath	Element	Type	Format
1	country	../name	Node	String	
2	capital	../capital	Node	String	
3	language	name	Node	String	
4	isofficial	isofficial	Attribute	String	
5	percentage	percentage	Node	Number	

4- Clique em Preview rows para visualizar uma prévia dos dados extraídos.



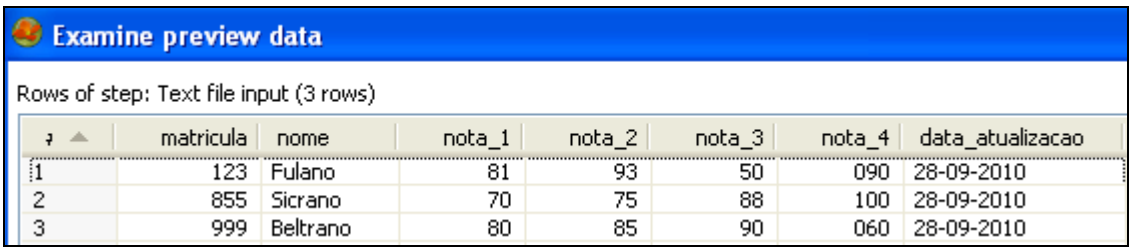
#	country	capital	language	isofficial	percentage
1	Afghanistan	Kabul	Pashto	T	524,0
2	Afghanistan	Kabul	Dari	T	321,0
3	Afghanistan	Kabul	Uzbek	F	88,0
4	Afghanistan	Kabul	Turkmenian	F	19,0
5	Afghanistan	Kabul	Balochi	F	9,0
6	Albania	Tirana	Albaniana	T	979,0
7	Albania	Tirana	Greek	F	18,0
8	Albania	Tirana	Macedonian	F	1,0
9	Algeria	Alger	Arabic	T	86,0
10	Algeria	Alger	Berberi	F	14,0
11	American Samoa	Fagatogo	Samoan	T	906,0
12	American Samoa	Fagatogo	English	T	31,0
13	American Samoa	Fagatogo	Tongan	F	31,0
14	Andorra	Andorra la Vella	Spanish	F	446,0

5- Salve a transformação e edite o Step Excel output para gravar o resultado em uma planilha.
6- Rode a transformação e veja os dados dos países em formato de planilha.

Exercício 11 – Calculando valores de novos campos

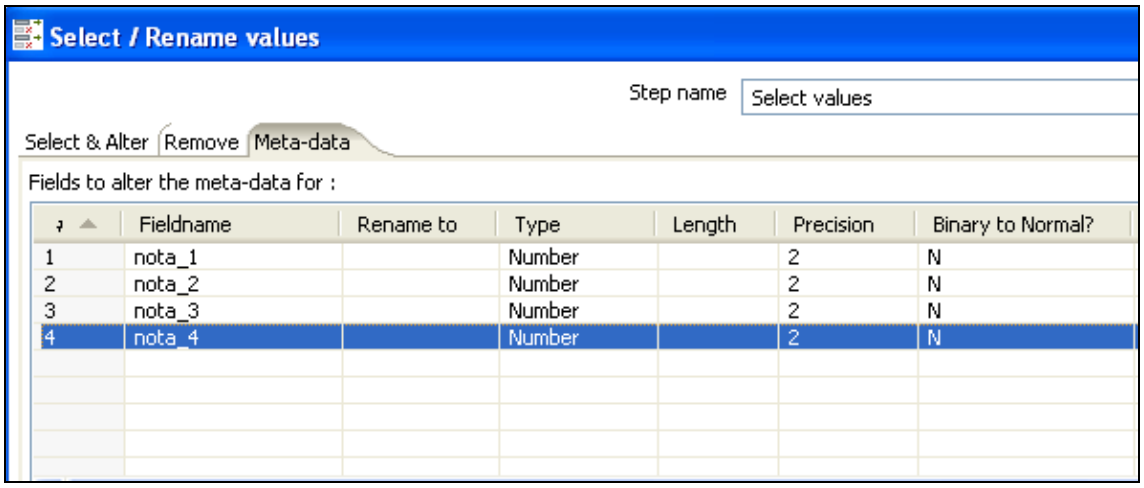
1 – Para executar esse exercício, localize o arquivo *notas-fulano.txt* na pasta *Bases* do material do curso. Esse arquivo descreve as notas de 3 alunos. Queremos converter as notas da escala de 0 a 10 para uma escala de 0 a 5 e calcular a média para cada aluno.

2 – Crie uma nova transformação e adicione um step *Text file input*. Edite o step e informe a localização do arquivo, conforme mostrado nos exercícios anteriores. Na aba *Fields*, clique em *Obtém campos* para recuperar os cabeçalhos dos campos e faça um preview dos dados.



#	matricula	nome	nota_1	nota_2	nota_3	nota_4	data_atualizacao
1	123	Fulano	81	93	50	090	28-09-2010
2	855	Sicrano	70	75	88	100	28-09-2010
3	999	Beltrano	80	85	90	060	28-09-2010

3 – Adicione um step *Select values*. Crie um hop ligando o *Text file input* ao *Select values*. Edite esse último step com os parâmetros mostrados na figura abaixo. Esse step é necessário para formatar os dados lidos no arquivo.



#	Fieldname	Rename to	Type	Length	Precision	Binary to Normal?
1	nota_1		Number		2	N
2	nota_2		Number		2	N
3	nota_3		Number		2	N
4	nota_4		Number		2	N

4 – Da categoria *Transform*, crie um step do tipo *Calculator*. Crie um hop ligando o *Select values* a esse novo step. Edite o step de acordo com a figura abaixo.

Calculadora

Nome do Step: Calculator

Campos:

#	Novo campo	Cálculo	Campo A	Campo B	Campo C	Tipo do valor	Tamanho	Precisão	Remove
1	fator_divisao	Define o valor constante A para o campo	20			Integer			N
2	nota_1_nova	A / B	nota_1	fator_divisao		Number		2	N
3	nota_2_nova	A / B	nota_2	fator_divisao		Number		2	N
4	nota_3_nova	A / B	nota_3	fator_divisao		Number		2	N
5	nota_4_nova	A / B	nota_4	fator_divisao		Number		2	N
6	numero_notas	Define o valor constante A para o campo	4			Integer			N
7	aux1	A + B	nota_1_nova	nota_2_nova		Number		2	N
8	aux2	A + B	nota_3_nova	nota_4_nova		Number		2	N
9	total	A + B	aux1	aux2		Number		2	N
10	media	A / B	total	numero_notas		Number		2	N

O que fizemos aqui? Basicamente, definimos um fator de conversão das notas (*fator_divisao*). Depois, dividimos as quatro notas pelo fator, criando quatro novos campos (*nota_1_nova*, ...). Em seguida, utilizamos quatro campos auxiliares para fazer o somatório das notas (*numero_notas*, *aux1*, *aux2* e *total*). Por fim, calculamos a média e armazenamos o resultado em um novo campo (*media*). Faça um preview na transformação e você deverá ver os novos campos.

Examine preview data

Rows of step: Calculator (3 rows)

#	matricula	nome	nota_1	nota_2	nota_3	nota_4	data_atualizacao	fator_divisao	nota_1_nova	nota_2_nova	nota_3_nova
1	123	Fulano	81,00	93,00	50,00	90,00	28-09-2010	20	4,05	4,65	2,5
2	855	Sicrano	70,00	75,00	88,00	100,00	28-09-2010	20	3,5	3,75	4,4
3	999	Beltrano	80,00	85,00	90,00	60,00	28-09-2010	20	4,0	4,25	4,5

5 – Crie um novo step *Select values* e hop ligando o *Calculator* ao novo step. Selecione e altere os campos que serão enviados para o step de carregamento, de acordo com a figura abaixo.

Select / Rename values

Step name: Select values 2

Select & Alter | Remove | Meta-data

Fields :

#	Fieldname	Rename to	Length	Precision
1	media			
2	matricula			
3	nome			
4	nota_1_nova	nota1		
5	nota_2_nova	nota2		
6	nota_3_nova	nota3		
7	nota_4_nova	nota4		
8	data_atualizacao			

6 – Crie um step *Dummy* apenas para visualizar os dados em modo de preview (se preferir, salve o resultado em um arquivo texto ou planilha). O preview dos dados deverá ser semelhante ao da figura abaixo. Salve a transformação.

Examine preview data

Rows of step: Dummy Plugin (3 rows)

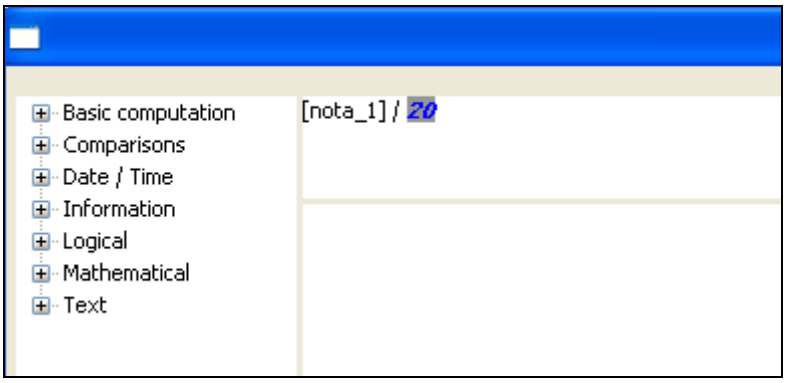
#	media	matricula	nome	nota1	nota2	nota3	nota4	data_atualizacao
1	3,925	123	Fulano	4,05	4,65	2,5	4,5	28-09-2010
2	4,1625	855	Sicrano	3,5	3,75	4,4	5,0	28-09-2010
3	3,9375	999	Beltrano	4,0	4,25	4,5	3,0	28-09-2010

Exercício 12 – Calculando valores dos campos – versão com step *Formula*.

1 – Vamos explorar uma nova forma de converter as notas e calcular as médias do exemplo anterior. Abra a transformação salva anteriormente e exclua o step *Calculator*.

2 – Na categoria *Scripting*, crie um novo step do tipo *Formula* e posicione-o no lugar ocupado pelo *Calculator*. Recrie os hops que foram excluídos quando você apagou o *Calculator*.

3- Edite o step *Formula*. Em *New Field*, adicione um campo chamado *nota1*. Clique na célula da coluna *Formula*, ao lado do campo que você acabou de criar. Será exibida uma janela de diálogo (despreze a mensagem de erro). Na parte superior da janela, digite `[nota_1]/20`. Clique em Ok e volte para a grade de inclusão dos campos.



4 – Repita o processo para os demais campos, de acordo com a figura abaixo.

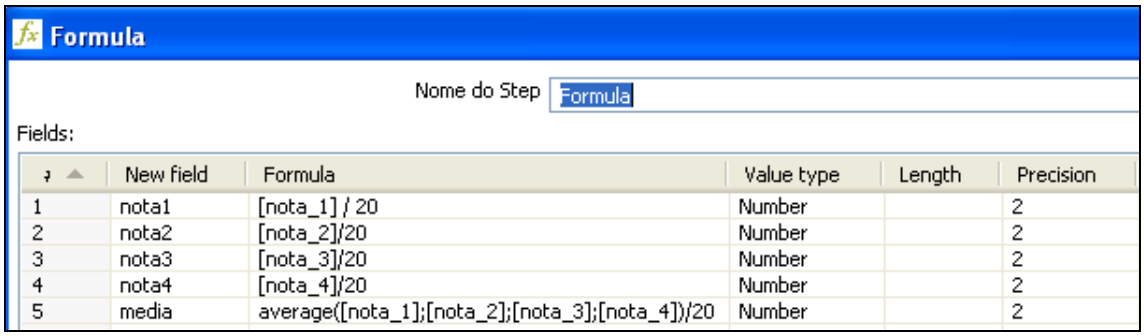
Formula

Nome do Step: Formula

Fields:

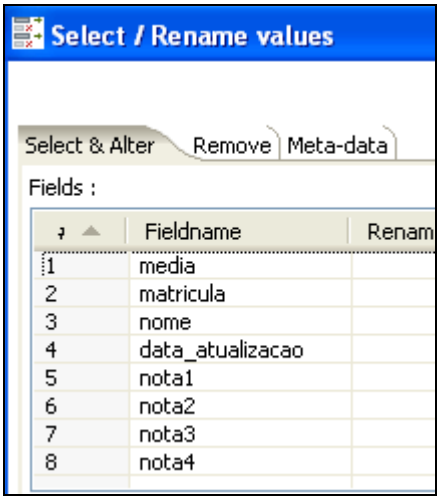
#	New field	Formula	Value type	Length
1	nota1	[nota_1] / 20	Number	
2	nota2	[nota_2]/20	Number	
3	nota3	[nota_3]/20	Number	
4	nota4	[nota_4]/20	Number	

5 – Crie um quinto campo chamado media e digite a fórmula **average**`([nota_1];[nota_2];[nota_3];[nota_4])/20`.



#	New field	Formula	Value type	Length	Precision
1	nota1	[nota_1] / 20	Number		2
2	nota2	[nota_2]/20	Number		2
3	nota3	[nota_3]/20	Number		2
4	nota4	[nota_4]/20	Number		2
5	media	average([nota_1];[nota_2];[nota_3];[nota_4])/20	Number		2

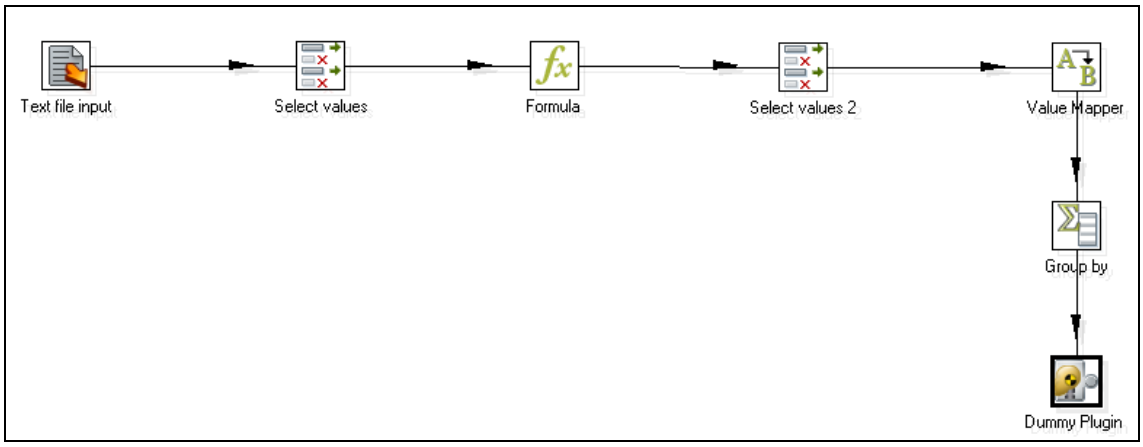
6 – Edite o segundo step Select values para selecionar os campos que serão armazenados, conforme a figura abaixo. Teste a transformação em preview. Salve a transformação.



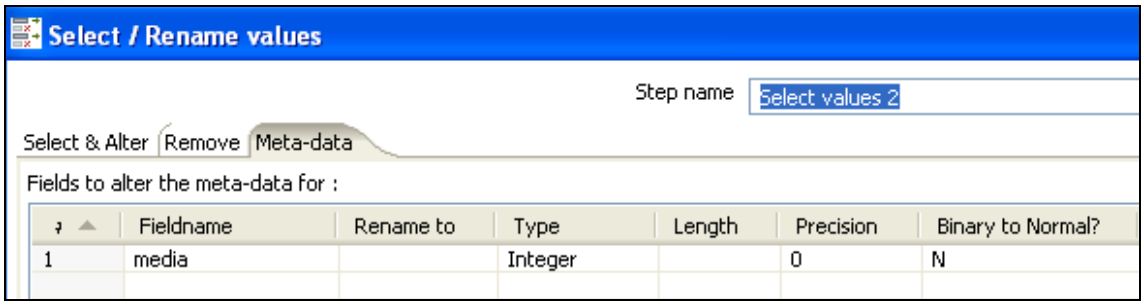
#	Fieldname	Renam
1	media	
2	matricula	
3	nome	
4	data_atualizacao	
5	nota1	
6	nota2	
7	nota3	
8	nota4	

Exercício 13 – Criando agregações em grupos de linhas.

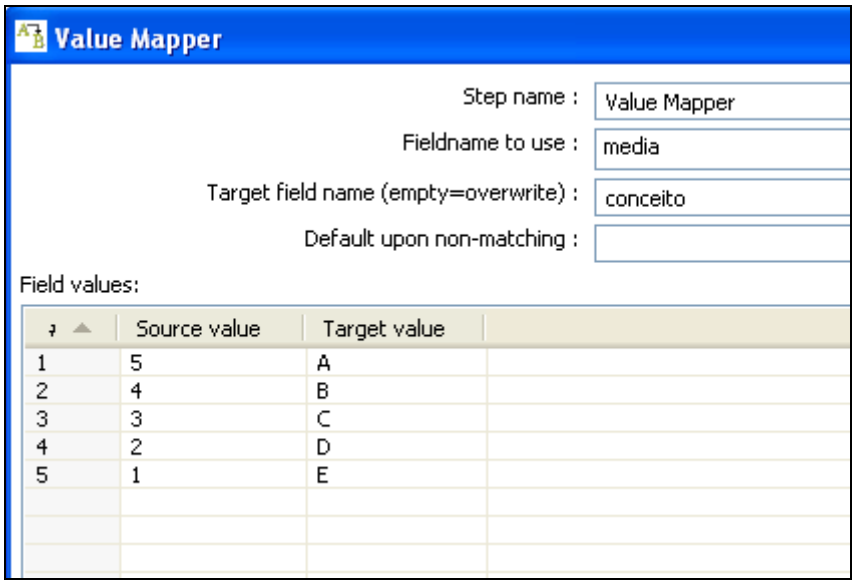
1 – Abra a transformação do exercício anterior. Crie um step *Value Mapper*, da categoria *Transform* e ligue o último *Select values* com ele. Em seguida, crie um step *Group by*, da categoria *Statistics*.



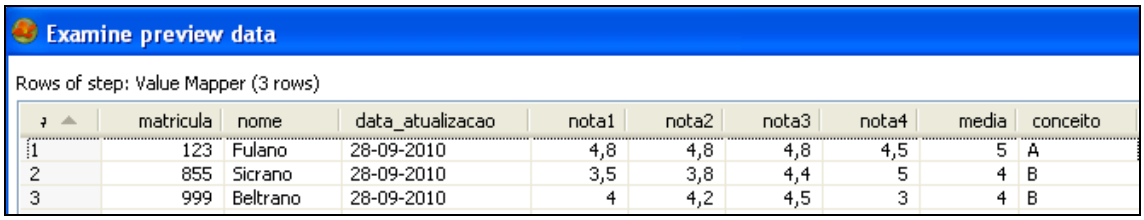
2 – Altere o segundo *Select values*, na aba *Meta-data*, para arredondar o valor do campo *media*.



3 – Edite o step Value Mapper, com os seguintes parâmetros da figura abaixo:

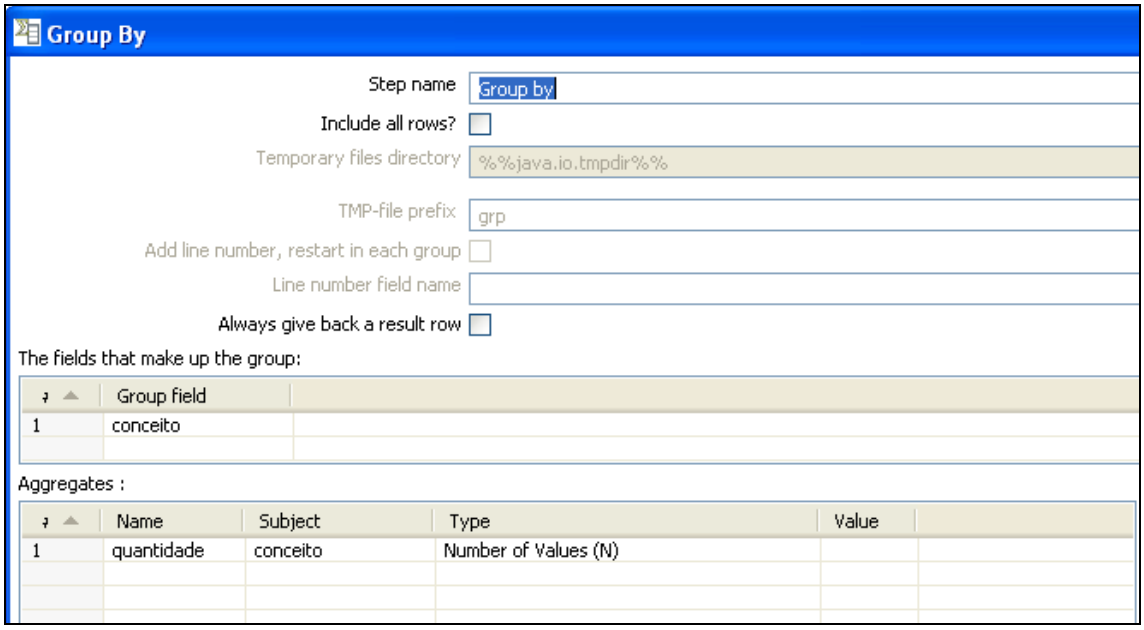


Faça um preview dos dados e veja se o step mapeou o valor arredondado da média para um conceito, variando entre A e E.



#	matricula	nome	data_atualizacao	nota1	nota2	nota3	nota4	media	conceito
1	123	Fulano	28-09-2010	4,8	4,8	4,8	4,5	5	A
2	855	Sicrano	28-09-2010	3,5	3,8	4,4	5	4	B
3	999	Beltrano	28-09-2010	4	4,2	4,5	3	4	B

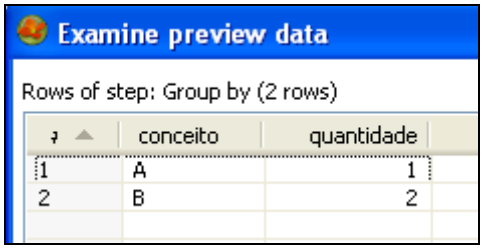
4 – Edite o step Group by de acordo com a figura abaixo.



#	Group field
1	conceito

#	Name	Subject	Type	Value
1	quantidade	conceito	Number of Values (N)	

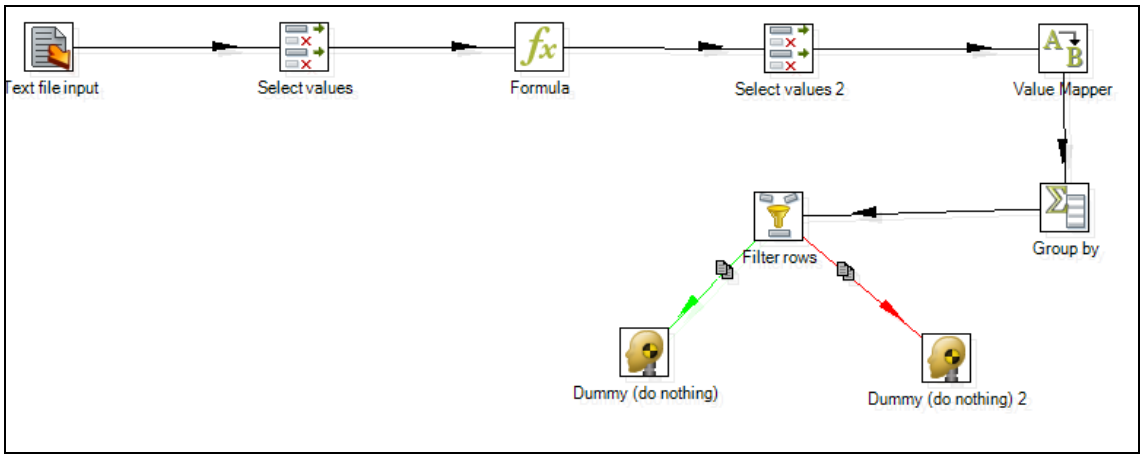
Esse step irá agregar as linhas a partir campo *conceito* e exibir o número de ocorrências de cada agrupamento (fique à vontade para testar outras funções de agregação desse step). Salve a transformação e testa-a no preview.



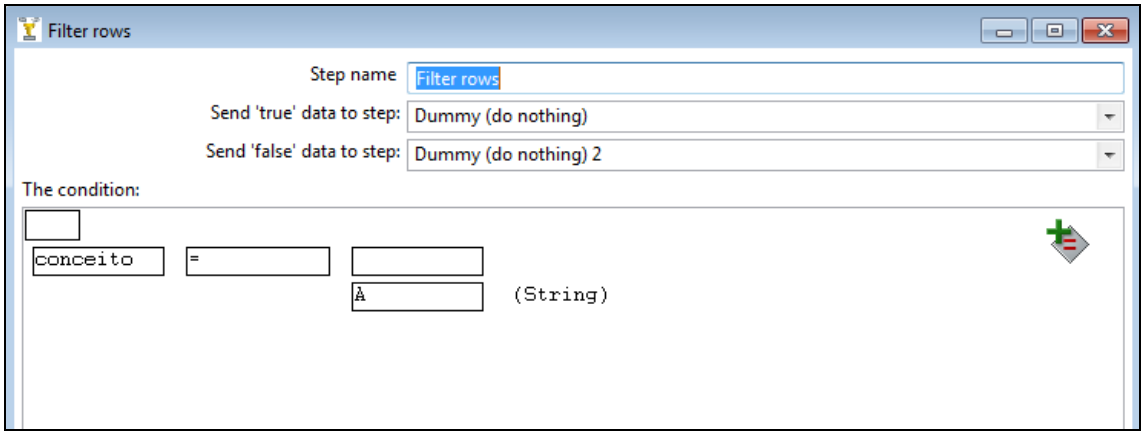
#	conceito	quantidade
1	A	1
2	B	2

Exercício 14 – Filtrando linhas de um dataset

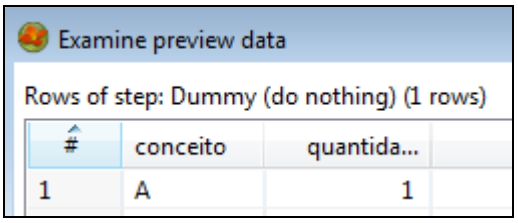
1 – Abra a transformação anterior e crie um step do tipo *Filter rows*, da categoria *Flow*. Crie um segundo step *Dummy* e crie os hops de acordo com a figura abaixo.

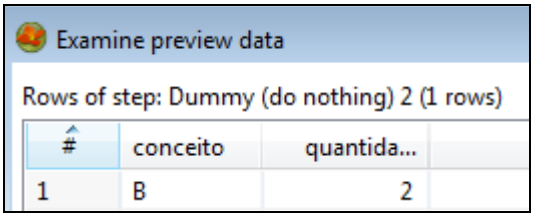


2 – Edite o step *Filter rows*, com os parâmetros da figura abaixo.



3 – Salve a transformação e faça um preview em cada step *Dummy* da transformação. Note que o fluxo enviado para cada step varia de acordo com a condição informada no *Filter rows*.





Exercício 15 – Exemplo de cópia do stream do dataset

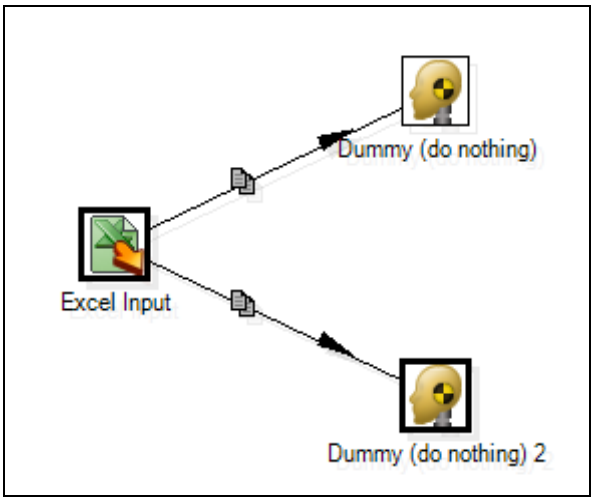
1 – Abra uma nova transformação e crie um step *Excel input*. Acesse o arquivo *areas.xls*, disponível na pasta bases do material do curso. Obtenha a planilha (aba *Sheets*) e recupere os campos (aba *Fields*, com o botão *Get fields...*). Dê um preview para visualizar os dados carregados.

Examine preview data

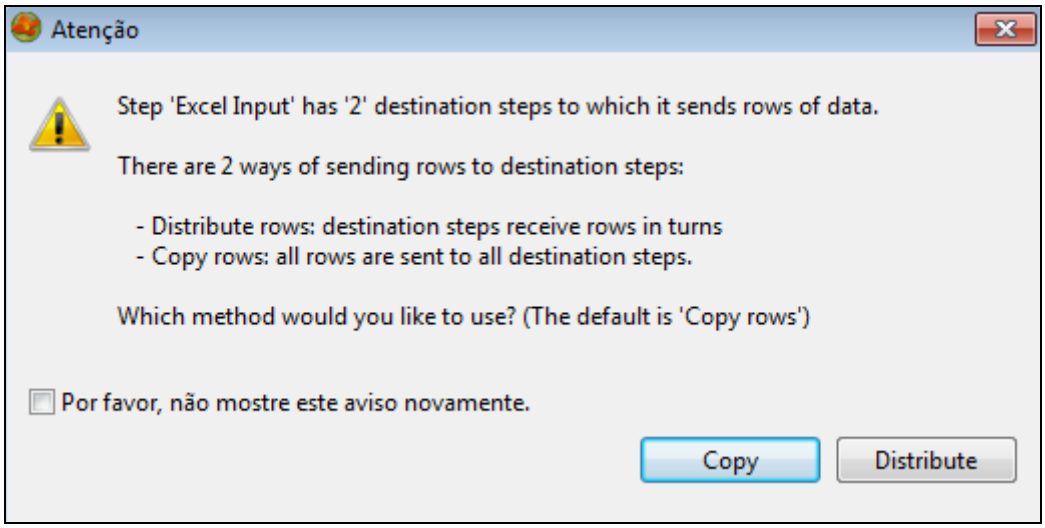
Rows of step: Excel Input (45 rows)

#	area	cod_area
1	Administração, Ciências Contábeis e Turis...	27
2	Antropologia / Arqueologia	35
3	Arquitetura e Urbanismo	29
4	Artes / Música	11
5	Astronomia / Física	3
6	Ciência da Computação	2
7	Ciência de Alimentos	25
8	Ciência Política e Relações Internacionais	39
9	Ciências Agrárias	42
10	Ciências Biológicas I	6
11	Ciências Biológicas II	8
12	Ciências Biológicas III	9
13	Ciências Sociais Aplicadas I	31

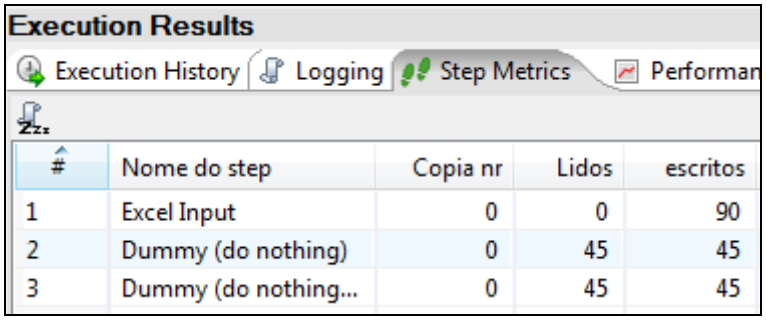
2 – Crie dois steps do tipo Dummy e dois hops, saindo simultaneamente do step Excel input e conectados com cada Dummy, de acordo com a figura abaixo.



Ao aparecer a janela de diálogo, responda que você quer realizar uma cópia.



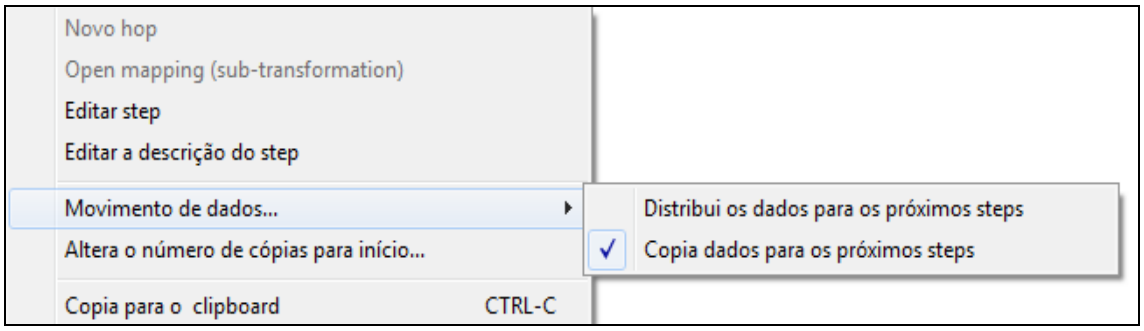
3 – Execute a transformação e observe a quantidade de registros lidos por cada step *Dummy*.



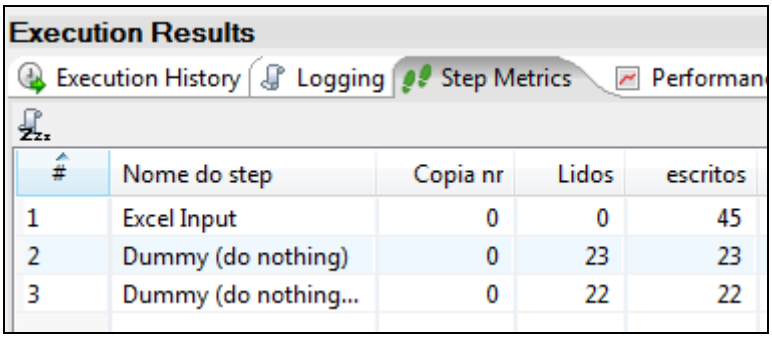
#	Nome do step	Copia nr	Lidos	escritos
1	Excel Input	0	0	90
2	Dummy (do nothing)	0	45	45
3	Dummy (do nothing...	0	45	45

Exercício 16 – Exemplo de distribuição do stream do dataset

1 – Abra a transformação do exercício anterior, clique com o botão direito em cima do step *Excel input* e escolha a opção *Movimento dos dados -> Distribui os dados para os próximos steps*.



2 – Execute a transformação e veja o número de registros lidos por cada step *Dummy*.



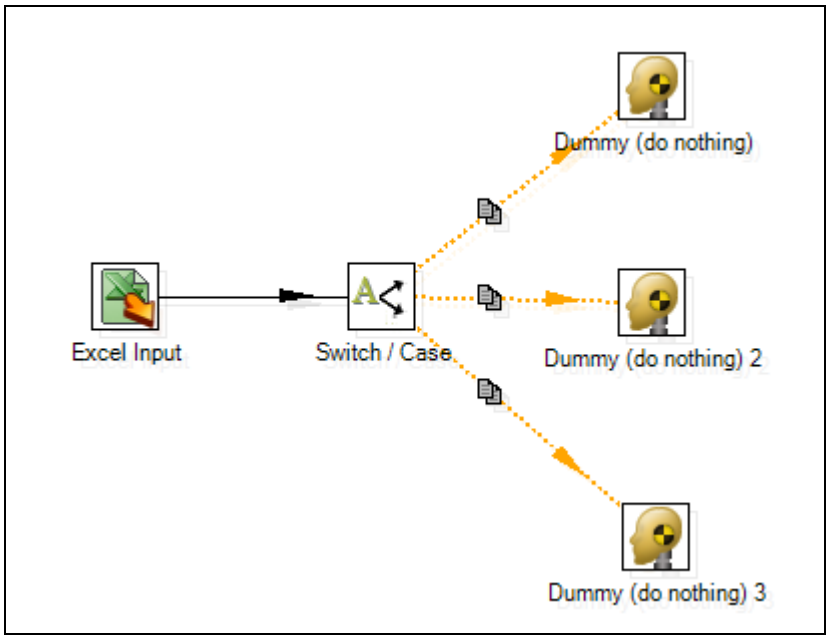
Execution Results

Execution History | Logging | Step Metrics | Performan

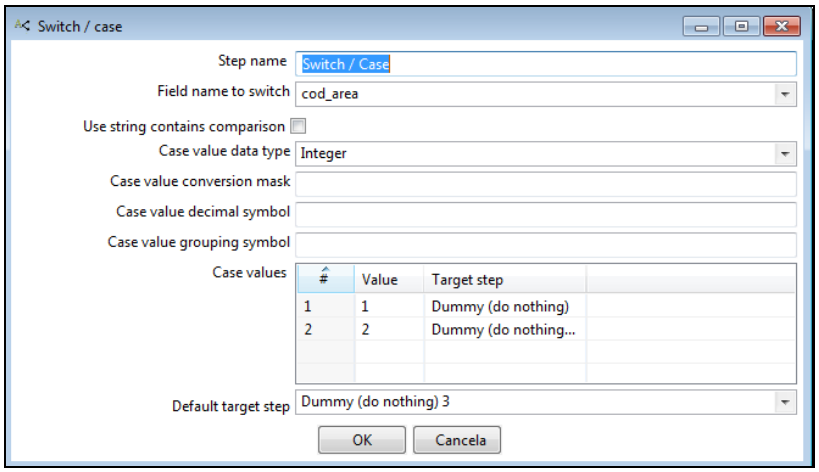
#	Nome do step	Copia nr	Lidos	escritos
1	Excel Input	0	0	45
2	Dummy (do nothing)	0	23	23
3	Dummy (do nothing...	0	22	22

Exercício 17 – Exemplo de distribuição do stream usando o step *Switch/Case*.

1 – Abra a transformação do exercício anterior e acrescente um step *Switch/Case*, da categoria *Flow* e um terceiro step *Dummy*, conforme a figura abaixo.



2 – Edite o step Switch/case, de acordo com as configurações abaixo.



Switch / case

Step name: Switch / Case

Field name to switch: cod_area

Use string contains comparison

Case value data type: Integer

Case value conversion mask

Case value decimal symbol

Case value grouping symbol

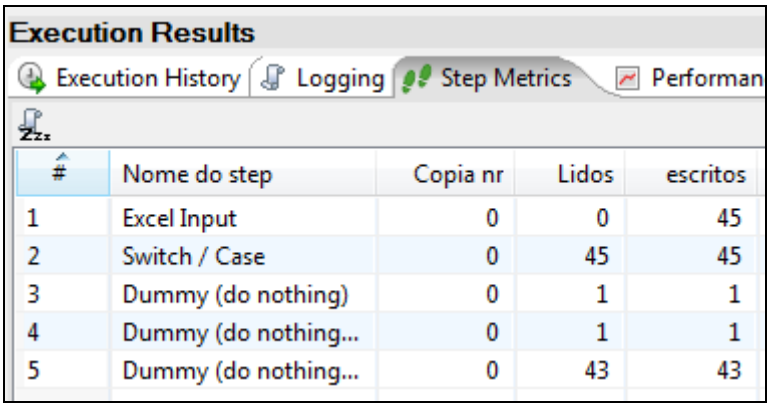
Case values

#	Value	Target step
1	1	Dummy (do nothing)
2	2	Dummy (do nothing...

Default target step: Dummy (do nothing) 3

OK | Cancela

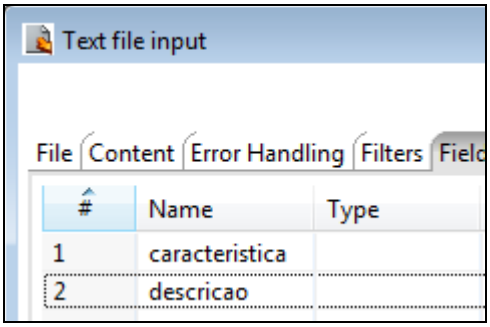
3 – Salve a transformação e execute-a, observando a quantidade de registros lidos por cada step *Dummy*. Caso queira conferir os registros que foram lidos em cada step *Dummy*, faça um preview e observe como o step *Switch/Case* realizou o filtro baseado no valor do campo *cod_area*.



#	Nome do step	Copia nr	Lidos	escritos
1	Excel Input	0	0	45
2	Switch / Case	0	45	45
3	Dummy (do nothing)	0	1	1
4	Dummy (do nothing...	0	1	1
5	Dummy (do nothing...	0	43	43

Exercício 18 – Fazendo conversões no *rowset*.

1 – Para realizar esse exercício, precisaremos do arquivo *movies.txt*, localizado na pasta *bases* do material distribuído. Abra uma nova transformação e crie um step *Text File input*. Acesse o arquivo *movies.txt*, adicione à lista de arquivos selecionados. Na aba *Contents* coloque *:* (dois pontos) como separador. Desmarque as opções *Header* e *No empty rows*. Na aba *Fields*, crie dois campos: *caracteristica* e *descricao*. Obtenha uma preview dos dados e veja os valores carregados.



#	Name	Type
1	caracteristica	
2	descricao	

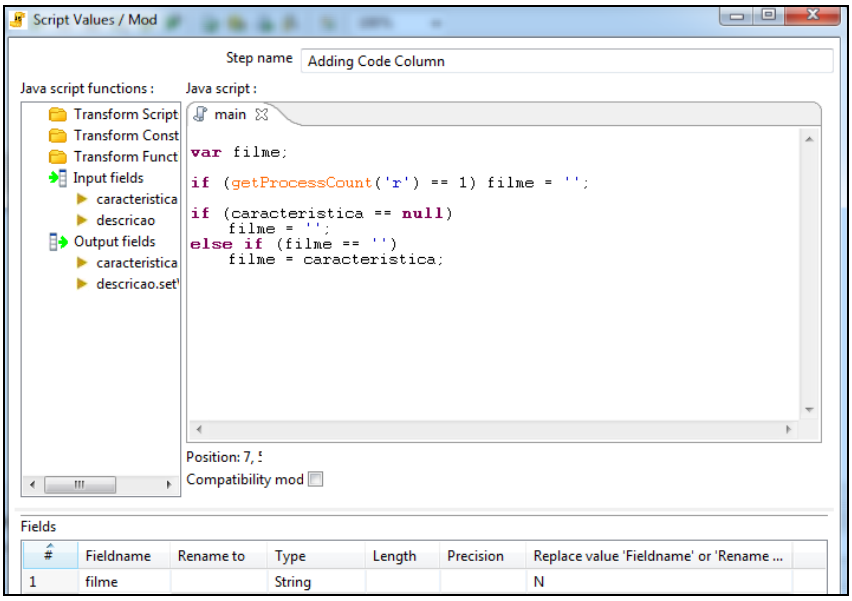
Examine preview data

Rows of step: Text file input (583 rows)

#	caracteristica	descricao
1	Persepolis	
2	Year	2007
3	Genre	Animation | Comedy | Drama | History
4	Directed by	Vincent Paronnaud, Marjane Satrapi
5	Script	Vincent Paronnaud, Marjane Satrapi
6	Music	Olivier Bernet
7	Cast	Chiara Mastroianni, Catherine Deneuve, Danielle Darrieux
8		
9	Trois couleurs - Rouge	
10	Year	1994
11	Genre	Drama
12	Directed by	Krzysztof Kieslowski
13	Cast	Irène Jacob, Jean-Louis Trintignant, Frédérique Feder, Jean-Pierre L
14		
15	Les Misérables	
16	Year	1933
17	Genre	Drama | History
18	Directed by	Raymond Bernard

Note que as características variam em quantidade para cada filme.

2 – Nesse exercício, vamos utilizar alguns steps voltados para auxiliar o trabalho do projetista. Inicialmente, vamos criar um step do tipo *Modified Javascript value*, da categoria *Scripting*. Esse step permite a criação de campos através de linhas de código em Javascript. Queremos criar uma coluna *Film*, no dataset lido do arquivo texto. Ligue o Text Input file a esse step com um hop e edite-o, de acordo com os parâmetros abaixo.



Esse código cria um campo *Film* e preenche o seu valor com o nome do filme. Dê um preview no step e verifique se o campo será preenchido corretamente.

Examine preview data

Rows of step: Adding Code Column (583 rows)

#	caracteristica	descricao	filme
1	Persepolis		Persepolis
2	Year	2007	Persepolis
3	Genre	Animation | Comedy | Drama | History	Persepolis
4	Directed by	Vincent Paronnaud, Marjane Satrapi	Persepolis
5	Script	Vincent Paronnaud, Marjane Satrapi	Persepolis
6	Music	Olivier Bernet	Persepolis
7	Cast	Chiara Mastroianni, Catherine Deneuve, Danielle Darrieux	Persepolis
8			
9	Trois couleurs - Rouge		Trois couleurs - Rouge
10	Year	1994	Trois couleurs - Rouge
11	Genre	Drama	Trois couleurs - Rouge
12	Directed by	Krzysztof Kieslowski	Trois couleurs - Rouge
13	Cast	Irène Jacob, Jean-Louis Trintignant, Frédérique Feder, Jean-Pierre Lorit, Samuel Le Bihan	Trois couleurs - Rouge
14			

3 – Em seguida, crie um step do tipo *Filter rows*. Nosso objetivo agora é filtrar todas as linhas que estão com o campo *descricao* nulo. Edite esse step com a seguinte condição:

Filter rows

Step name: Filter null rows

Send 'true' data to step:

Send 'false' data to step:

The condition:

descricao | IS NOT NULL | -

-

4 – O próximo passo é fazer a conversão dos dados de linhas para colunas. Para isso, vamos criar um step do tipo *Row denormalizer*, da categoria *Transform*. Edite esse step de acordo com a figura abaixo.

Denormaliser

Step name: 1 row x film

The key field: caracteristica

The fields that make up the grouping:

#	Group field
1	filme

Target fields:

#	Target fieldname	Value fieldname	Key value	Type	Format
1	Ano	descricao	Year	String	
2	Genêro	descricao	Genre	String	
3	Diretor	descricao	Directed by	String	
4	Atores	descricao	Cast	String	

A partir do campo *filme*, a linhas com o conteúdo A*no*, *Gênero*, *Diretor* e *Atores* são convertidas em colunas, com o valor do campo descrição. Dê um preview na transformação e veja os valores convertidos.

Examine preview data

Rows of step: 1 row x film (97 rows)

#	filme	Ano	Gênero	Diretor	Atores
1	Persepolis	2007	Animation | Comedy | Drama | History	Vincent Paronnaud, Marjane Satr...	Chiara Mastroianni, Catherine Deneuve, Danielle Darrieux
2	Trois couleurs - Rouge	1994	Drama	Krzysztof Kieslowski	Irène Jacob, Jean-Louis Trintignant, Frédérique Feder, Jean-P
3	Les Misérables	1933	Drama | History	Raymond Bernard	
4	Au revoir, les enfants	1987	Drama	Louis Malle	
5	La France	2007	Drama | Musical | Romance | War	Serge Bozon	Sylvie Testud, Pascal Greggory, Guillaume Verdier
6	L'Atalante	1934	Drama | Romance		Michel Simon, Dita Parlo, Jean Dasté

5 – O último passo é preencher as colunas com valores vazios com a string “n/a”. Para isso, vamos criar um step do tipo *If field value is null*, da categoria *Utility*. Edite o step de acordo com os parâmetros abaixo.

Replace null value

Step name: Defaults for null fields

Replace Null for all fields

Replace by value:

Mask (Date):

Select fields ☑

Select value type ☐

Value types

#	Type	Replace by value	Conversion mask (Date)

Fields

#	Field	Replace by value	Conversion mask (Date)
1	Gênero	n/a	
2	Diretor	n/a	

Salve a transformação e execute o seu preview. O resultado deverá ser semelhante ao da figura abaixo.

Examine preview data

Rows of step: Defaults for null fields (97 rows)

#	filme	Ano	Gênero	Diretor	Atores
1	Persepolis	2007	Animation | Comedy | Drama | History	Vincent Paronnaud, Marjane Satr...	Chiara Mastroianni, Catherine Deneuve, Danielle Darrie
2	Trois couleurs - Rouge	1994	Drama	Krzysztof Kieslowski	Irène Jacob, Jean-Louis Trintignant, Frédérique Feder, J
3	Les Misérables	1933	Drama | History	Raymond Bernard	
4	Au revoir, les enfants	1987	Drama	Louis Malle	
5	La France	2007	Drama | Musical | Romance | War	Serge Bozon	Sylvie Testud, Pascal Greggory, Guillaume Verdier
6	L'Atalante	1934	Drama | Romance	n/a	Michel Simon, Dita Parlo, Jean Dasté
7	La môme	2007	Biography | Drama | Music	Olivier Dahan	Marion Cotillard, Sylvie Testud, Pascal Greggory, Emm
8	MR 73	2008	Crime | Drama	n/a	Daniel Auteuil, Olivia Bonamy, Catherine Marchal, Fran
9	Manon Des Sources		Drama | Romance	Claude Berri	Yves Montand, Daniel Auteuil, Emmanuelle Béart
10	Un Coeur en Hiver	1992	Drama | Romance | Music	Claude Sautet	Daniel Auteuil, Emmanuelle Béart, André Dussollier
11	Les Voleurs	1996	Crime | Drama | Romance	André Téchiné	Catherine Deneuve, Daniel Auteuil, Laurence Côte, Ben
12	Caché	2005	n/a	Michael Haneke	Daniel Auteuil, Juliette Binoche, Maurice Bénichou
13	Jean de Florette	1986	Historical drama	Claude Berri	Yves Montand, Gérard Depardieu, Daniel Auteuil
14	Le Ballon rouge	1956	Fantasy | Comedy | Drama	n/a	

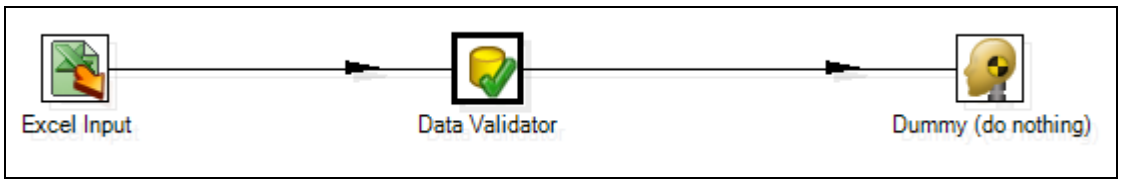
Exercício 19 – Validando dados do rowset.

1 – Para esse exercício vamos utilizar o arquivo *capes.xls*, localizado na pasta *bases* do material do curso. Crie um step *Excel input* e leia o arquivo, conforme mostrado nos exercícios anteriores.

2 – Em seguida, vamos criar um step para validar os campos do dataset lido. Queremos impedir que um determinado campo nulo seja repassado para os steps de carregamento. Crie um step do tipo *Data Validator*, da categoria *Validation*. Edite o step e clique no botão New validation. Dê um nome para a validação e edite os seus campos, de acordo com os parâmetros abaixo:

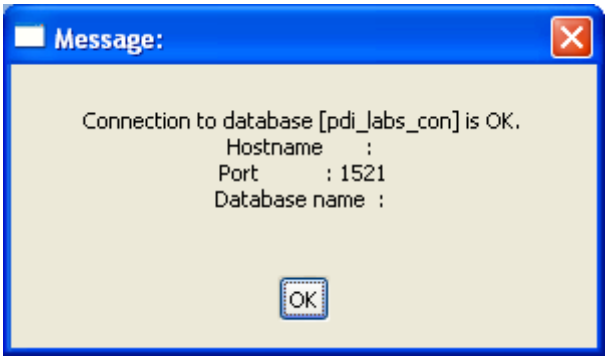
- Marque o campo *Report all erros, not only first*.
- Em *Name of field to validate*, escolha o campo *Escore_3*.
- No bloco *Data*, desmarque a opção *Null allowed?*

Salve a transformação e execute-a. Verifique que a transformação irá parar quando o primeiro valor nulo do campo *Escore_3* for encontrado.

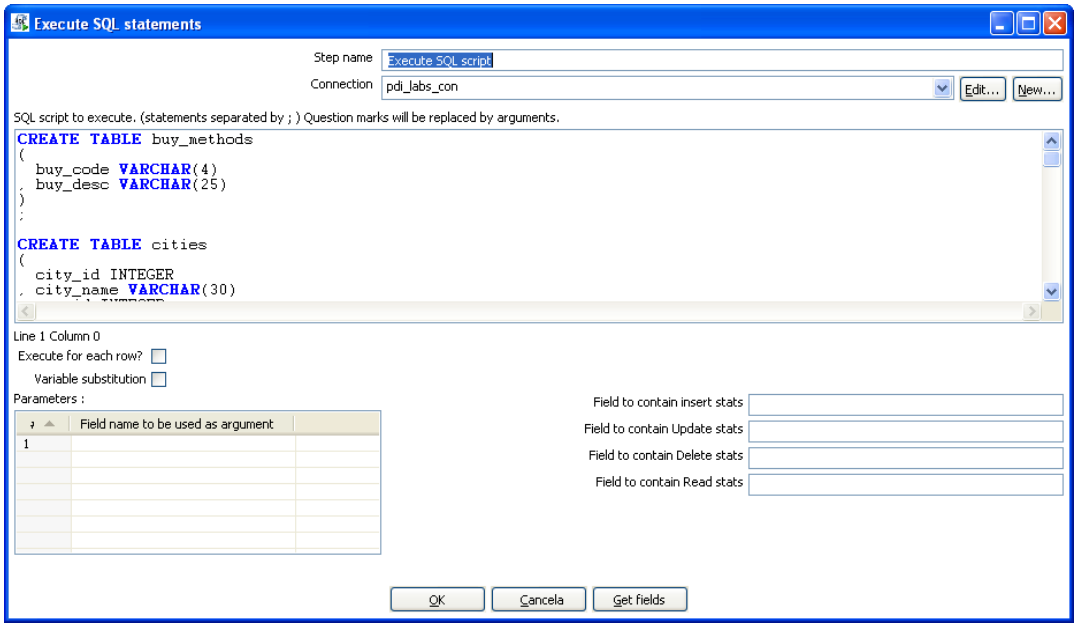


Exercício 20 – Criando as tabelas de um banco de dados

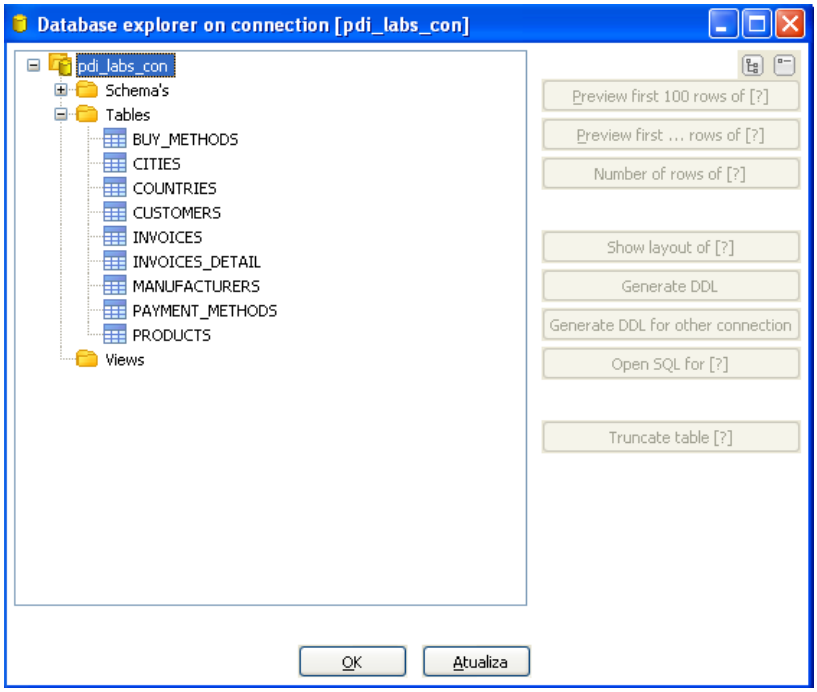
1 – Para esse exercício, vamos utilizar o arquivo *scripts-js.sql*, localizados na pasta script do material do curso. Esses scripts criam as tabelas que armazenam os dados de uma loja online de vendas de jogos. Antes de iniciar, verifique se a conexão *pdi_labs_con* (criada no Exercício 3) está disponível. Acesse a aba *View* e expanda a árvore de conexões. Edite a conexão e pressione o botão *Test*.



2 – Crie uma nova transformação e adicione um step do tipo *Execute SQL script*, da categoria *Scripting*. Edite o step, escolhendo a conexão *pdi_labs_con* no campo *Connection*. Em seguida, abra o arquivo *scripts-js.sql* em um editor de texto, copie e cole o seu conteúdo no campo *SQLscript to execute*. Salve a transformação e execute.

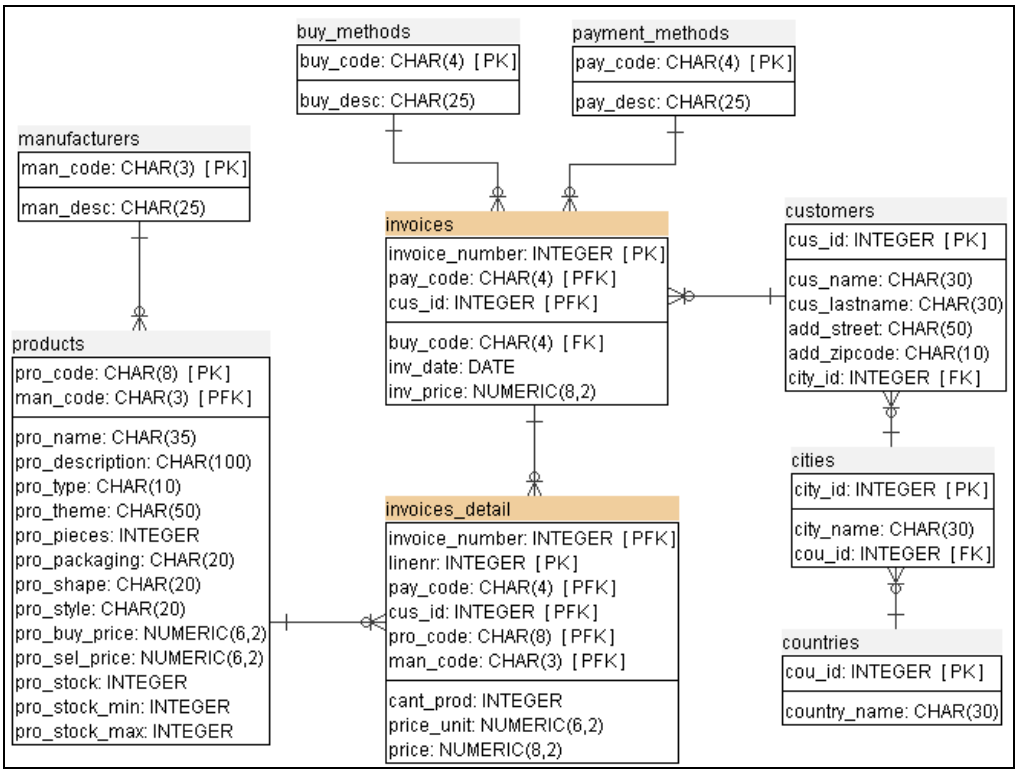


3 – Vamos confirmar se as tabelas foram realmente criadas. Na opção *View*, expanda a árvore de conexões e clique com o botão direito em cima da conexão *pdi_labs_con*, escolhendo a opção *Explorar*. Caso as tabelas tenham sido criadas corretamente, você poderá ver a lista expandindo a opção *Tables*.



A base de dados do exemplo possui 9 tabelas, de acordo com uma breve descrição de suas características abaixo:

Tabela	Descrição
manufacturers	Informações sobre os fabricantes dos produtos
products	Produtos que estão à venda na loja, como jogos, quebra-cabeça, acessórios, etc.
buy_methods	Contém informações sobre as modalidades de compra (física, por telefone, pela internet, etc.)
payment_methods	Contém informações sobre as modalidades de pagamento (dinheiro, cartão, cheque)
countries	Uma lista de países
cities	Uma lista de cidades
customers	Os clientes cadastrados na loja
Invoices/invoices_detail	Cabeçalho e detalhe das informações de faturas



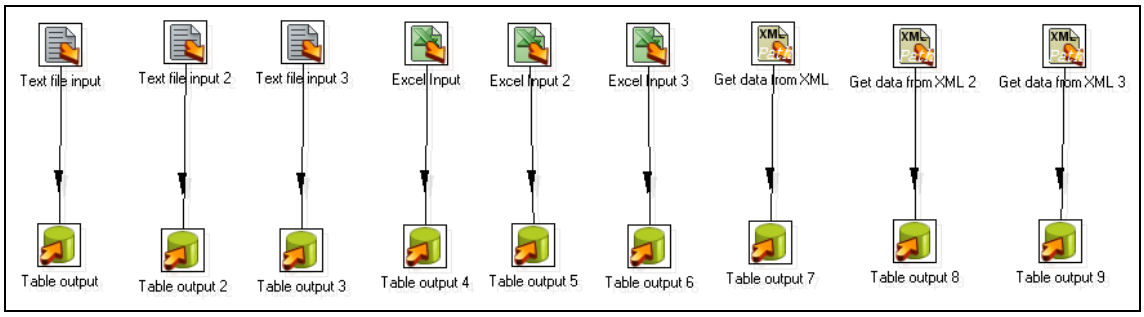
Exercício 21 – Carregando as tabelas através de transformações

1 – Vamos agora carregar os dados nas tabelas do banco a partir de um conjunto de arquivos localizados na pasta /bases/banco do material distribuído. Os arquivos que você deverá usar estão listados abaixo:

- BUY_METHODS.csv
- CITIES.csv
- COUNTRIES.csv
- CUSTOMERS.xls
- INVOICES.xls
- INVOICES_DETAIL.xls
- MANUFACTURERS.xml
- PAYMENT_METHODS.xml
- PRODUCTS.xml

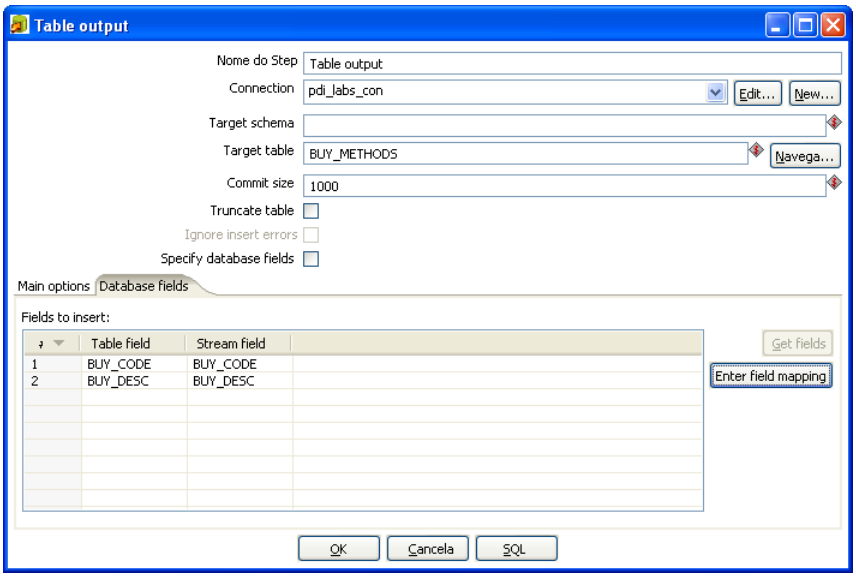
2 – Crie uma nova transformação e adicione vários steps para ler cada arquivo de entrada. Você terá que criar 3 *Text File input*, 3 *Excel input* e 3 *Get data from XML*. Conforme mostrado nos exercícios anteriores, para cada step de entrada, localize o arquivo e adicione-o à lista de arquivos selecionados. Configure o conteúdo na aba *Content* e obtenha os campos na aba *Fields*.

3 – Em seguida, crie 9 steps do tipo *Table output*, que irão armazenar os dados nas tabelas criadas no exercício anterior.



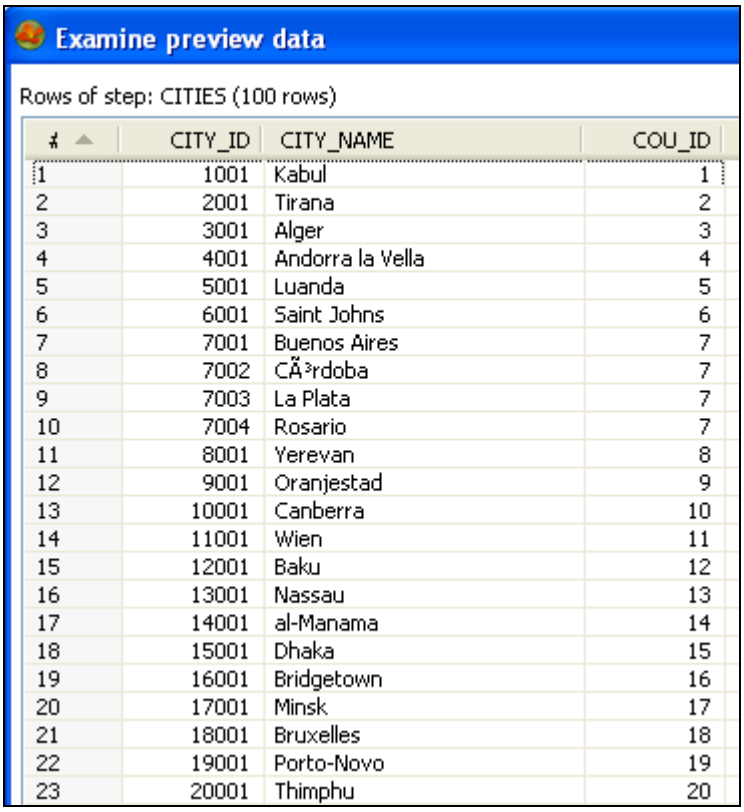
4 – Edite cada Table output da seguinte forma:

- *Connection*: escolha *pdi_labs_con*.
- *Target table*: uma das tabelas criadas no execício anterior. Clique no botão *Navegar* para escolher uma tabela da conexão. (ex.: *BUY_METHODS*)
- Aba *Database Fields*:
    - Clique no botão *Enter Field mapping* e escolha os mapeamentos adequados



- Salve a transformação e repita esse passo para cada step *Table output*.

5 – Para verificar se as tabelas foram realmente carregadas, clique na opção *View*, expanda a árvore de conexões e clique com o botão direito em cima da conexão *pdi_labs_con*, escolhendo a opção *Explorar*. Escolha uma tabela e clique na opção *Preview First 100 rows of <nome_da_tabela>.*



Examine preview data

Rows of step: CITIES (100 rows)

#	CITY_ID	CITY_NAME	COU_ID
1	1001	Kabul	1
2	2001	Tirana	2
3	3001	Alger	3
4	4001	Andorra la Vella	4
5	5001	Luanda	5
6	6001	Saint Johns	6
7	7001	Buenos Aires	7
8	7002	CÃ³rdoba	7
9	7003	La Plata	7
10	7004	Rosario	7
11	8001	Yerevan	8
12	9001	Oranjestad	9
13	10001	Canberra	10
14	11001	Wien	11
15	12001	Baku	12
16	13001	Nassau	13
17	14001	al-Manama	14
18	15001	Dhaka	15
19	16001	Bridgetown	16
20	17001	Minsk	17
21	18001	Bruxelles	18
22	19001	Porto-Novo	19
23	20001	Thimphu	20