

cin.ufpe.br



Centro de Informática

U • F • P • E



UNIVERSIDADE FEDERAL DE PERNAMBUCO

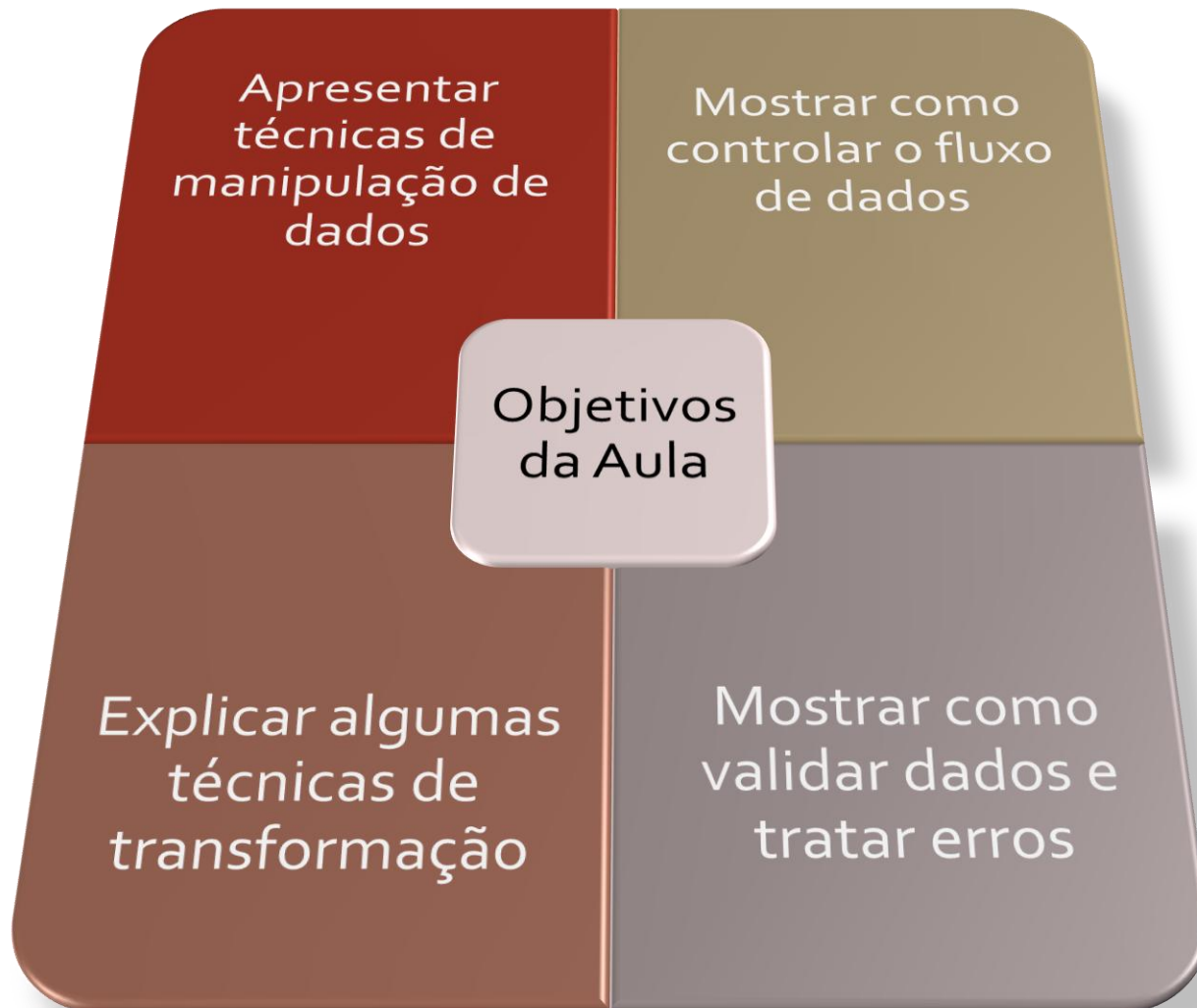


PENTAHO DATA INTEGRATION

SEMANA 2

Jarley Nóbrega – jpn@cin.ufpe.br

Pentaho Data Integration



|| Agenda

Manipulação de dados no PDI

Controlando o fluxo de dados

Transformações no *rowset*

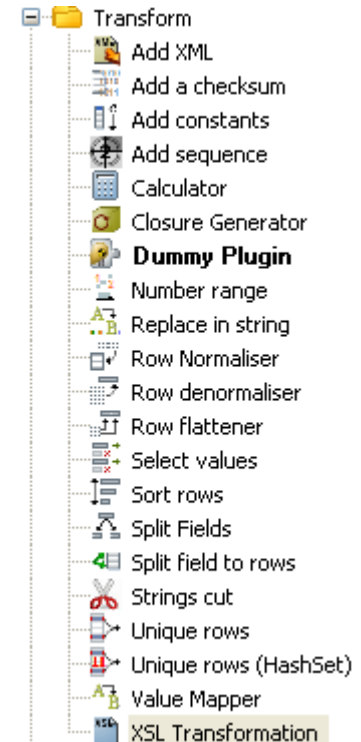
Tratamento de erros e validação de dados



MANIPULAÇÃO DE DADOS NO PDI

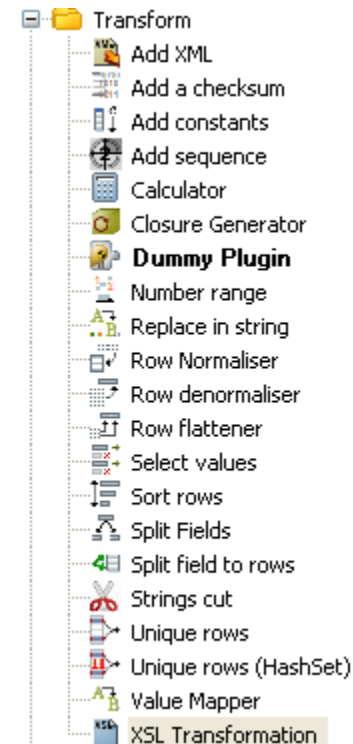
Manipulação básica de dados no PDI

- Conjunto de steps para transformação
 - Categoria *Transform*
 - Criação de novos campos
 - Uso de expressões
 - Adição de constantes
 - Cálculo de valores
 - Conversão de formatos
 - Correspondência de valores



Steps de Manipulação de dados

Step	Descrição
Calculator	Cria novos campos através de cálculos no <i>stream</i> .
Split fields	Divide um campo em dois ou mais através de um separador.
Add constants	Adiciona uma ou mais constantes ao <i>stream</i> .
Replace in string	Substitui todas as ocorrências em uma string por um texto.
Number range	Cria um novo campo baseado em uma faixa de valores numéricos
Value Mapper	Cria uma correspondência entre valores de um campo



Manipulação básica de dados no PDI

Dúvidas sobre o funcionamento de um tipo específico de step?

<http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+v3.2.+Steps>

|| Exercícios 11, 12 e 13

- Manipulando um conjunto de dados com os steps de transformação
- Criando grupos de linhas a partir de uma transformação

Manipulação básica de dados no PDI

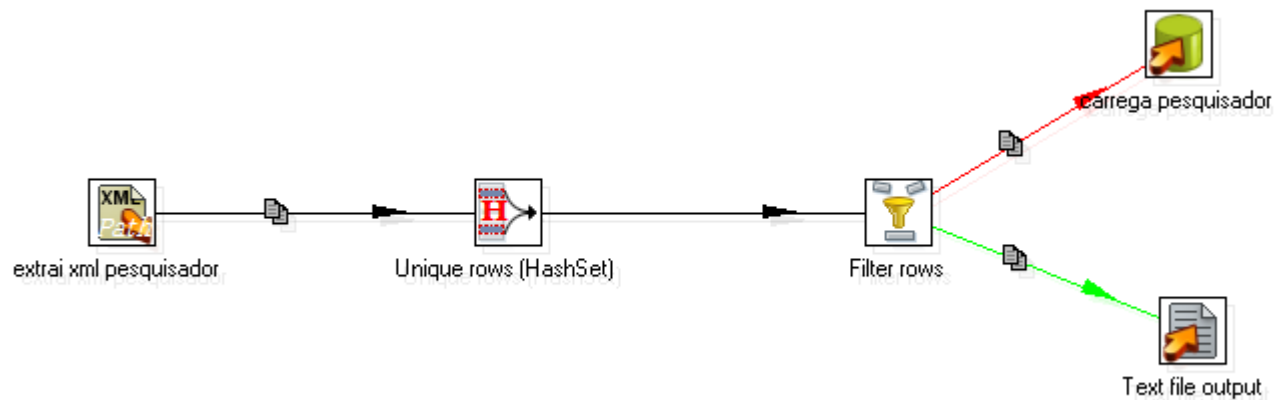
- Filtrando informações
 - Como descartar linhas de dados sob certas condições?
 - Step *Filter rows* (categoria Flow)



Filter rows

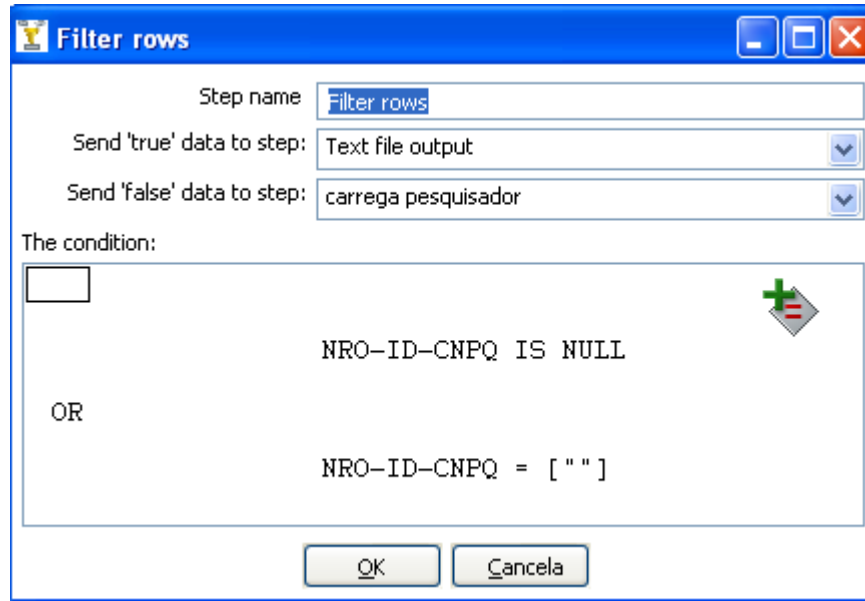
Manipulação básica de dados no PDI

- Step *Filter rows* (categoria Flow)
 - ▣ Checa a condição para cada rowset
 - ▣ Apenas as linhas cuja condição são `true` serão enviadas no *hop* para o próximo *step*.
 - ▣ Possibilidade de fazer *if-then-else*



Manipulação básica de dados no PDI

- Exemplo de condição




Filter rows

Step name:

Send 'true' data to step:

Send 'false' data to step:

The condition:

☐ 

NRO-ID-CNPQ IS NULL

OR

NRO-ID-CNPQ = [" "]

|| Exercício 14

- Filtrando linhas de um dataset



CONTROLANDO O FLUXO DE DADOS



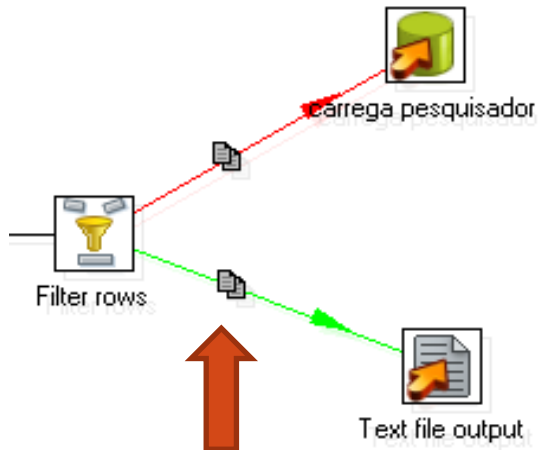
Controlando o fluxo de dados

Até agora...

- Fluxo simples de dados
 - Uma única direção
- Na maioria das vezes
 - Divisão de stream

Controlando o fluxo de dados

Como o PDI trata o fluxo de dados?



Cópia do Fluxo

- O dataset inteiro é copiado para os steps subsequentes

Distribuição do Fluxo

- O dataset é distribuído em partes iguais e enviado para os steps subsequentes

Controlando o fluxo de dados

- Como copiar ou distribuir um dataset a partir de uma condição?
 - 1ª opção: usar o *Filter rows*
 - Problema: "if's" aninhados
 - 2ª opção: usar o step *Switch/Case*

Switch / case

Step name: Switch / Case

Field name to switch: conceito_recomendado

Use string contains comparison: ☐

Case value data type: Integer

Case value conversion mask:

Case value decimal symbol:

Case value grouping symbol:

	Value	Target step
1	3	carrega_curso_nivel_3
2	4	carrega_curso_nivel_4
3	5	carrega_curso_nivel_5
4	6	carrega_curso_nivel_6
5	7	carrega_curso_nivel_7

Default target step:

OK Cancela

Steps de controle do fluxo de dados

Se você precisa...	Poderá usar...
Adicionar (append) um dataset em outro, sem importar a ordem.	qualquer step.
Adicionar (append) um dataset, usando uma ordem específica.	o step <i>Append streams</i> , da categoria <i>Flow</i> .
Fazer um merge com dois ou mais datasets, ordenados por um campo.	o step <i>Sorted merge</i> , da categoria <i>Joins</i> .
Fazer um merge com dois datasets, eliminando linhas duplicadas	o step <i>Merge rows (diff)</i> , da categoria <i>Joins</i> .

Exercícios 15, 16 e 17

- Copiando e distribuindo um dataset
- Usando condições para copiar e distribuir um dataset



TRANSFORMAÇÕES DO *ROWSET*



Transformações no *Rowset*

- Algumas transformações que podem ser feitas em cima de todo um *rowset*:
 - Converter linhas em colunas;
 - Converter colunas em linhas;
 - Operações em conjuntos de linhas;

Convertendo linhas em colunas

- Na maioria dos datasets cada linha pertence a um elemento diferente
- Em alguns casos, uma única linha não descreve completamente o elemento

```

...
{ Caché
  Year: 2005
  Director: Michael Haneke
  Cast: Daniel Auteuil, Juliette Binoche, Maurice Bénichou
{ Jean de Florette
  Year: 1986
  Genre: Historical drama
  Director: Claude Berri
  Produced by: Pierre Grunstein
  Cast: Yves Montand, Gérard Depardieu, Daniel Auteuil
{ Le Ballon rouge
  Year: 1956
  Genre: Fantasy | Comedy | Drama
...

```

Convertendo linhas em colunas

Solução do PDI: step *Row denormalizer*



- Converte linhas em colunas
- Exemplo:
 - ▣ Dataset final deverá ter uma única linha por filme

```
...
Caché
Year: 2005
Director: Michael Haneke
Cast: Daniel Auteuil, Juliette Binoche, Maurice Bénichou
Jean de Florette
Year: 1986
Genre: Historical drama
Director: Claude Berri
Produced by: Pierre Grunstein
Cast: Yves Montand, Gérard Depardieu, Daniel Auteuil
Le Ballon rouge
Year: 1956
Genre: Fantasy | Comedy | Drama
...
```

Convertendo linhas em colunas

Solução do PDI: step *Row denormalizer*



Row denormaliser



FILM	YEAR	GENRE	DIRECTOR	ACTORS

1 film
by row

...

Caché
Year: 2005
Director: Michael Haneke
Cast: Daniel Auteuil, Juliette Binoche, Maurice Bénichou
Jean de Florette
Year: 1986
Genre: Historical drama
Director: Claude Berri
Produced by: Pierre Grunstein
Cast: Yves Montand, Gérard Depardieu, Daniel Auteuil
Le Ballon rouge
Year: 1956
Genre: Fantasy | Comedy | Drama
...

Convertendo linhas em colunas

Solução do PDI: step *Row denormalizer*



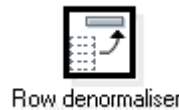
FILM	YEAR	GENRE	DIRECTOR	ACTORS

1 film
by row

- Regra geral para novos campos:
 - Se o valor para a chave de um campo for igual a "A", coloque o valor do campo "B" em um novo campo "C".

Convertendo linhas em colunas

Outra aplicação do step
Row denormalizer



- Agregação de dados
 - A partir de um dataset de entrada, gerar na saída um novo dataset com dados consolidados ou agregados.
 - Semelhança com ferramentas de *cross tab* (ex.: Pivot no Excel)

Steps de conversão em rowsets

Step	Descrição
Group By	Cria agregações em grupos de linhas usando Sum, Maximum, etc.
Univariate statistics	Computa estatísticas básicas em grupos de linhas
Split fields	Divide um campo em dois ou mais campos
Row normalizer	Transforma colunas em linhas
Row flattener	Faz um nivelamento nas linhas consecutivas
Sort rows	Ordena linhas através de uma chave
Split fields to rows	Divide um campo de string e cria uma nova linha para cada termo da divisão
Unique rows	Remove linhas duplicadas no dataset (precisa de ordenação prévia)

|| Exercício 18

- Fazendo conversões no *rowset*

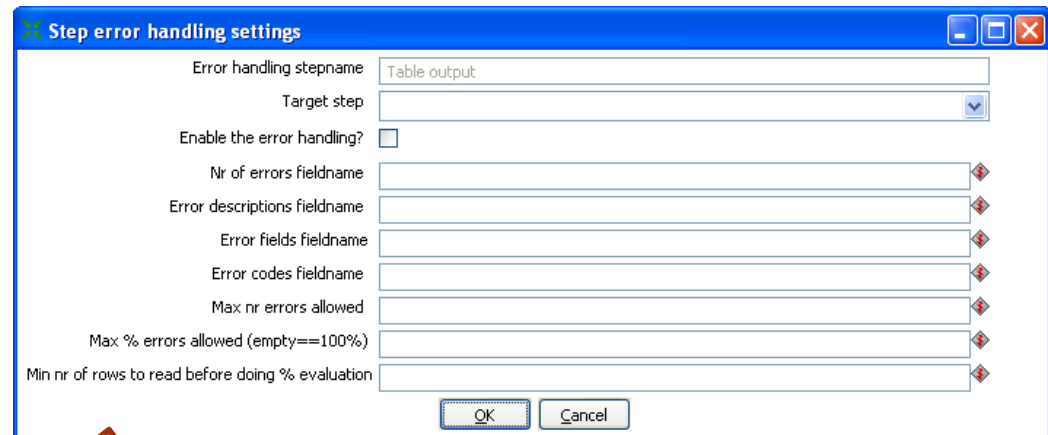
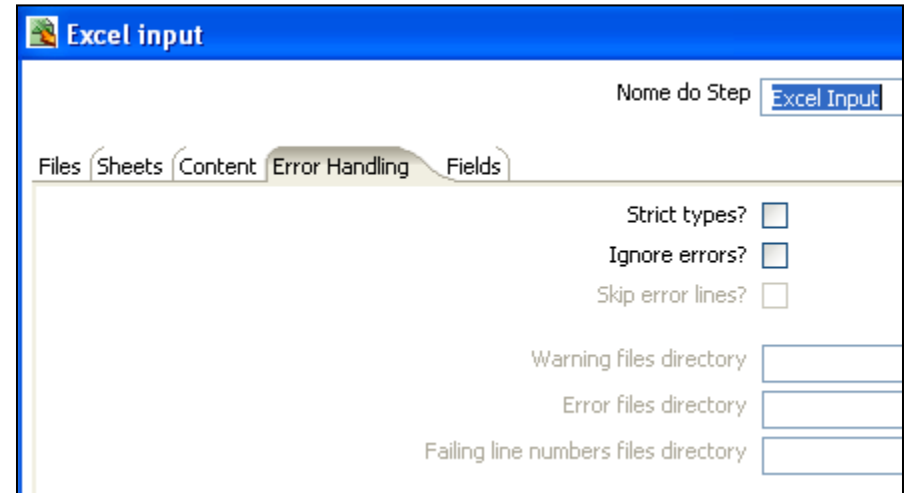
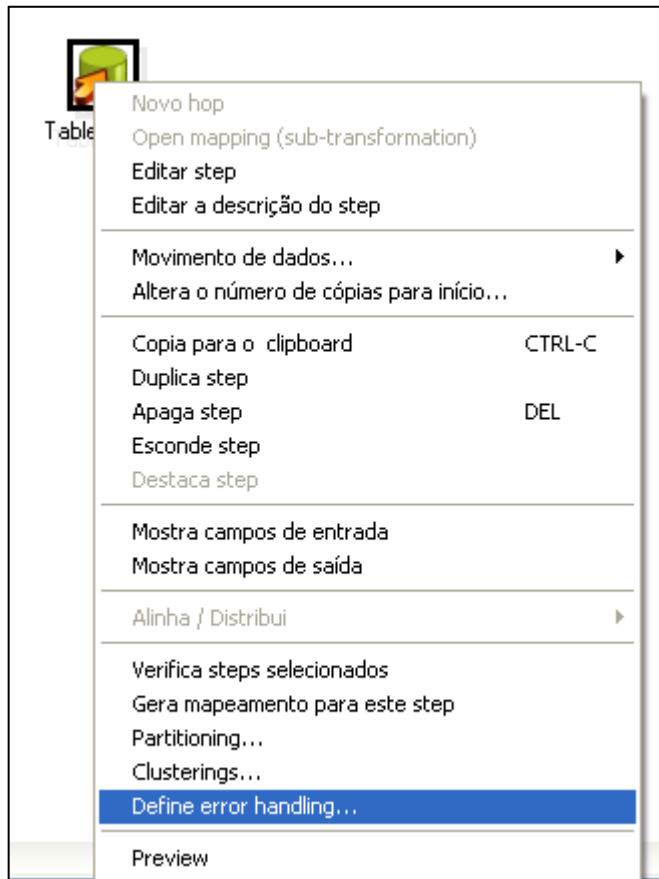


TRATAMENTO DE ERROS E VALIDAÇÃO DOS DADOS

Tratamento de Erros

- Até agora...
 - Erros encontrados nas transformações
 - Janela de *Logging*
- Tratamento de erros no PDI
 - Opção *Define Error handling* (disponíveis em alguns steps)
 - Opção *Error handling* (disponível na edição dos steps)

Tratamento de Erros



Tratamento de Erros - Configurações

Campo do Step	Descrição
Nr of errors fieldname	Nome do campo que irá armazenar o número de erros
Error fields fieldname	Nome do campo que registrará os campos onde ocorreram os erros
Error codes fieldname	Nome do campo que contém o código do erro
Error descriptions fieldname	Nome do campo que contém a descrição do erro

Step error handling settings

Error handling stepname: Table output

Target step: [dropdown]

Enable the error handling? ☐

Nr of errors fieldname: [text box]

Error descriptions fieldname: [text box]

Error fields fieldname: [text box]

Error codes fieldname: [text box]

Max nr errors allowed: [text box]

Max % errors allowed (empty==100%): [text box]

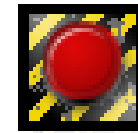
Min nr of rows to read before doing % evaluation: [text box]

OK Cancel

Tratamento de Erros

Como tratar?

- Detectando o erro e enviando as linhas com problemas para outro *stream*.
- Quando a quantidade de erros é grande? Quando os erros são críticos?
- Opção: usar o step *Abort*, da categoria *Flow*



Abort

Tratamento de Erros

Como personalizar um arquivo de log no PDI?

- Step *Write to log*, categoria *Utility*



Write to log

Write to log

Step name: Write to log

Log level: Basic logging

Print header: ☒

Fields

	Field
1	

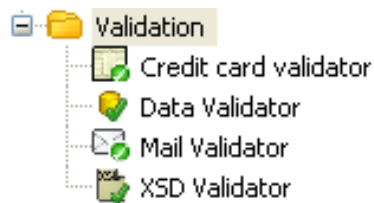
OK Obtem campos Cancela

Validação de dados



Validação de dados

Categoria de steps *Validation*



- Step *Data Validator*
 - Validação de tipos de dados
 - Validação do conteúdo dos dados



Validação de dados – step *Data Validator*

Validação desejada	Bloco de opções <i>Data</i>
Permitir (apenas) valores nulos	Null allowed? / Only null values allowed?
Tamanho de um campo está dentro de uma faixa de valores	Max string length / Min string length
Valor de um campo está dentro de uma faixa de valores	Maximum value / Minimum value
Campo selecionado atende a um padrão	Only numeric data expected, Expected start string, Expected end string, Regular expression expected to match
Campo selecionado não atende a um padrão	Not allowed start string, Not allowed end string, Regular expression not allowed to match
Campo selecionado é um dos valores permitidos em uma lista	Allowed values, Read allowed values from another step?

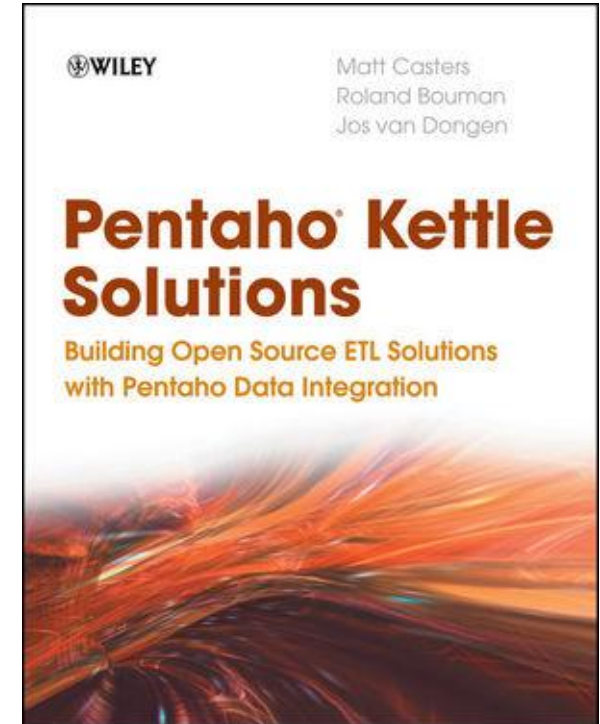
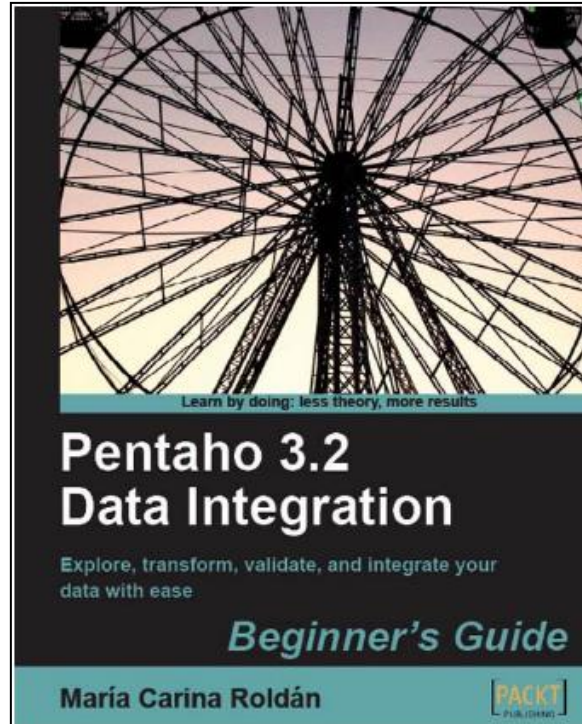
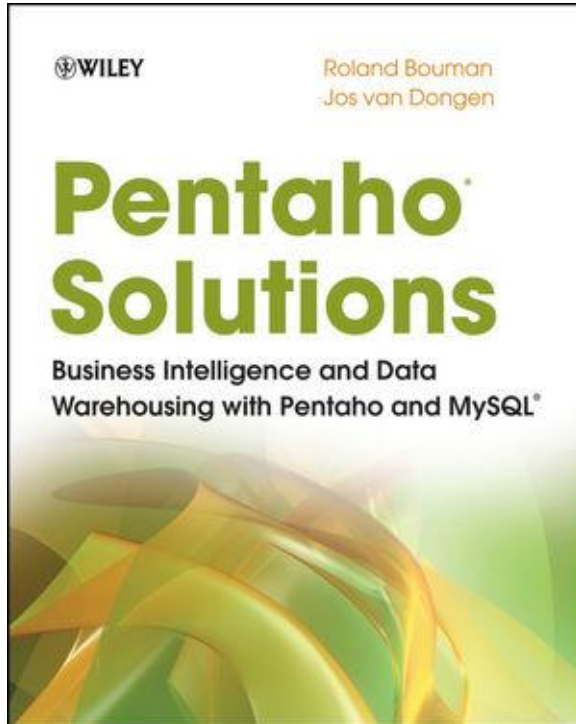
|| Exercício 19

- Validando o conteúdo dos dados

Resumo da Semana 2

- Steps de transformação
- Filtragem de dados
- Cópia e distribuição do *stream*
- Transformações no *rowset*
- Tratamento de erros
- Validação do tipo e conteúdo dos dados

Bibliografia



Site do PDI: <http://kettle.pentaho.com/>



Perguntas?