cin.ufpe.br

Centro de Informática

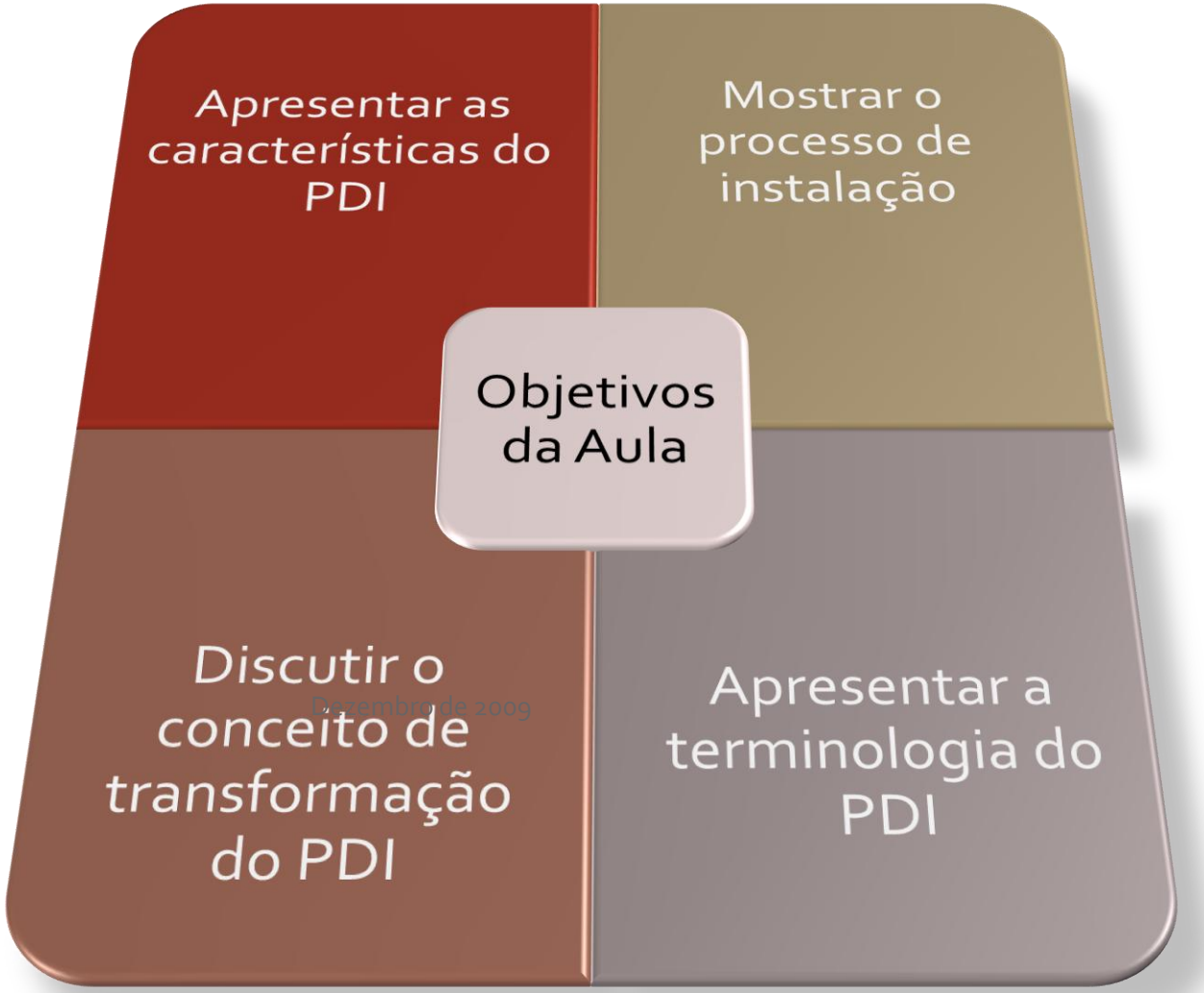
U · F · P · E

Universidade Federal de Pernambuco

PENTAHO DATA INTEGRATION
SEMANA 1

Jarley Nóbrega – jpn@cin.ufpe.br

Pentaho Data Integration



Agenda

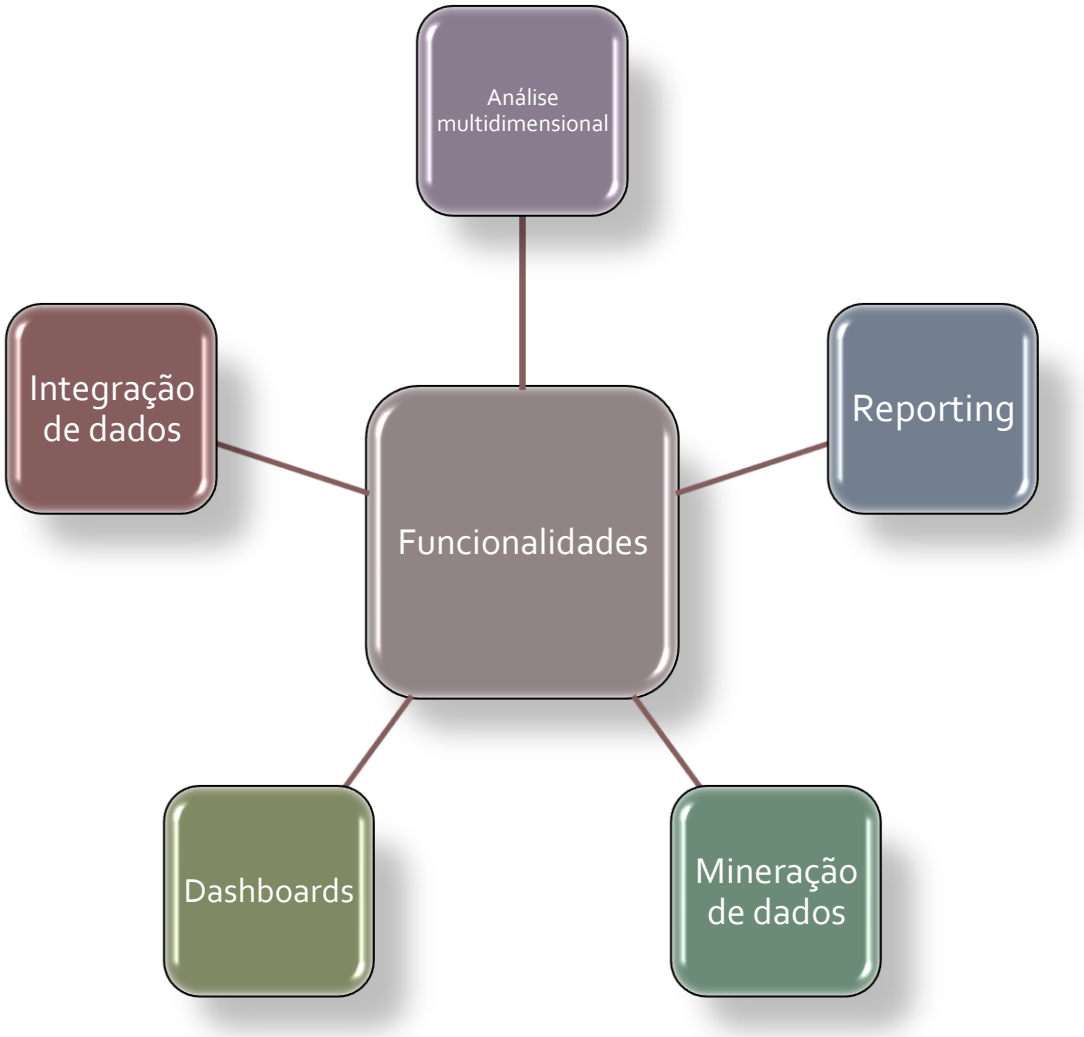
- O PDI e o Pentaho BI Suite
- Instalando o PDI
- Trabalhando com arquivos

INTRODUÇÃO AO PENTAHO BI SUITE

Pentaho BI Suite

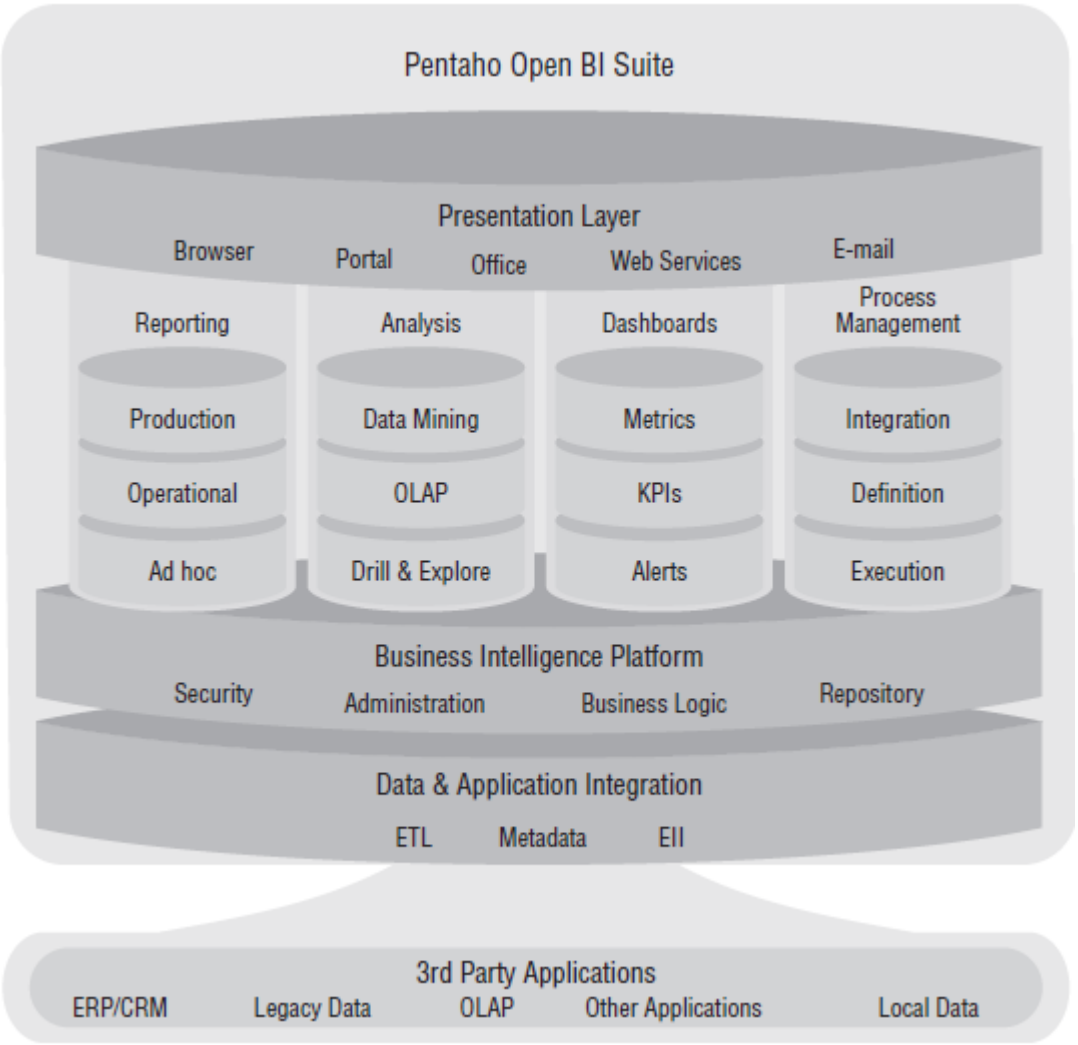
- Coleção de Aplicações de Software
  - Criação e *deployment* de soluções para tomada de decisão
  - Open source
    - Enterprise /Community Editions
  - http://www.pentaho.com

Pentaho BI Suite



Pentaho BI Suite

Arquitetura do Pentaho BI



Camadas da arquitetura do Pentaho BI Suite (Bouman and Dongen, 2009)

Pentaho BI Suite

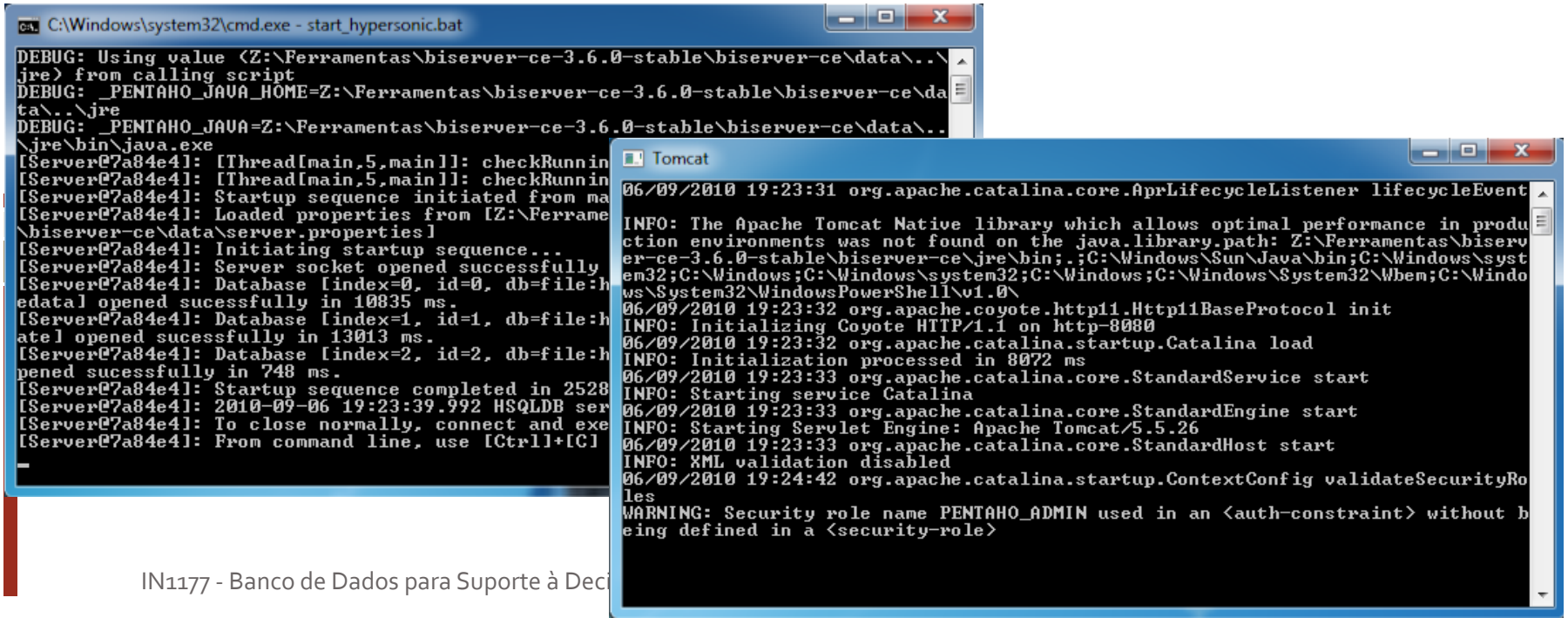
- Pentaho BI Platform demo
  - Instalação pré-configurada da plataforma Pentaho
  - Demonstração do uso de relatórios, cubos e dashboards
    - Base de dados Steel Wheels
- Download
  - http://sourceforge.net/projects/pentaho/files/
  - Pasta Business Intelligence Server: arquivo `biserver-ce-3.6.0.stable.zip` (~170MB)

Pentaho BI Suite

- Um pequeno roteiro para rodar o BI Server
  - Baixar e descompactar o arquivo
  - Certifique-se que existe uma JVM instalada
    - Verifique a variável de ambiente JAVA_HOME
    - Se estiver no Linux, dê acesso de gravação e leitura para a pasta do tomcat.
      - `sudo chmod 755 ./tomcat/*`

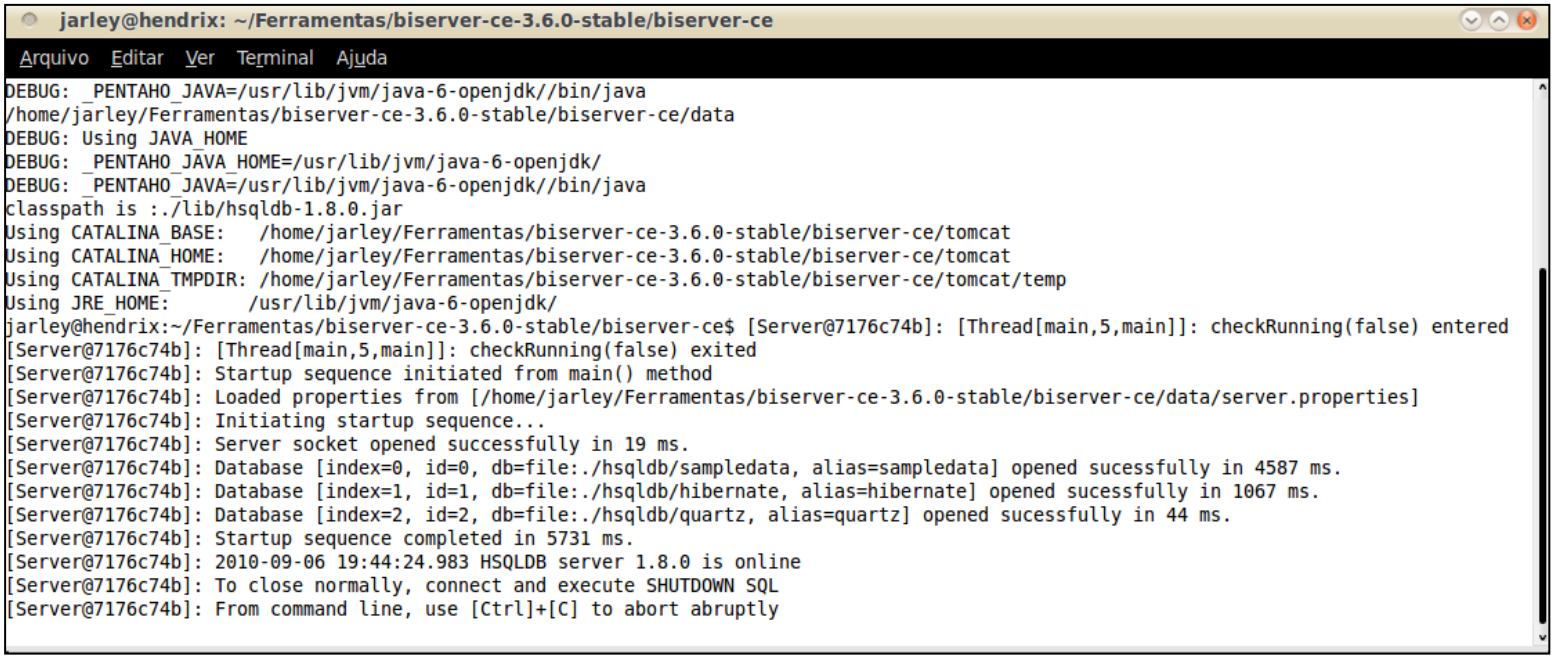
Pentaho BI Suite

- Um pequeno roteiro para rodar o BI Server
  - Inicie o BI Server
    - Windows: `<pasta>\bi-server-ce\start-pentaho.bat`



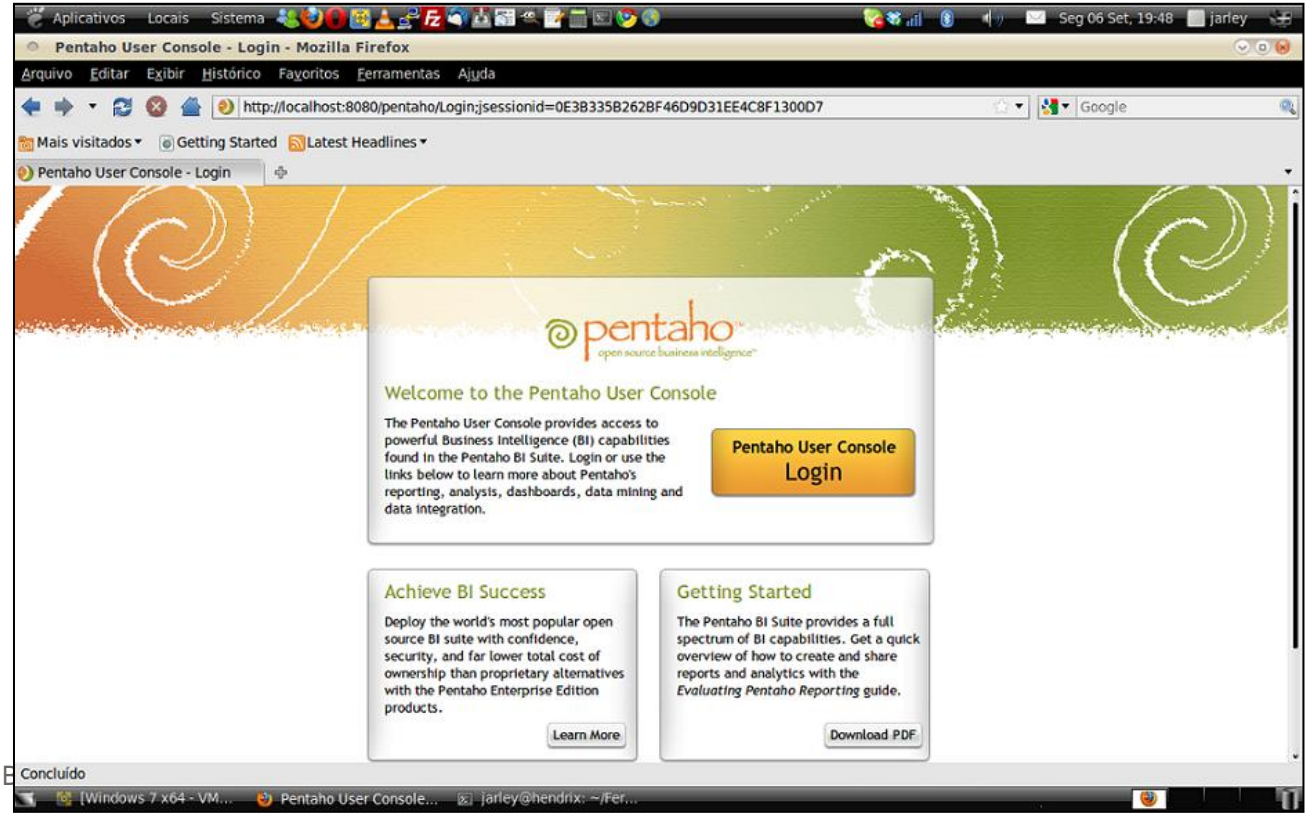
Pentaho BI Suite

- Um pequeno roteiro para rodar o BI Server
  - Inicie o BI Server
    - Linux: `<pasta>/bi-server-ce/sh ./start-pentaho.sh`



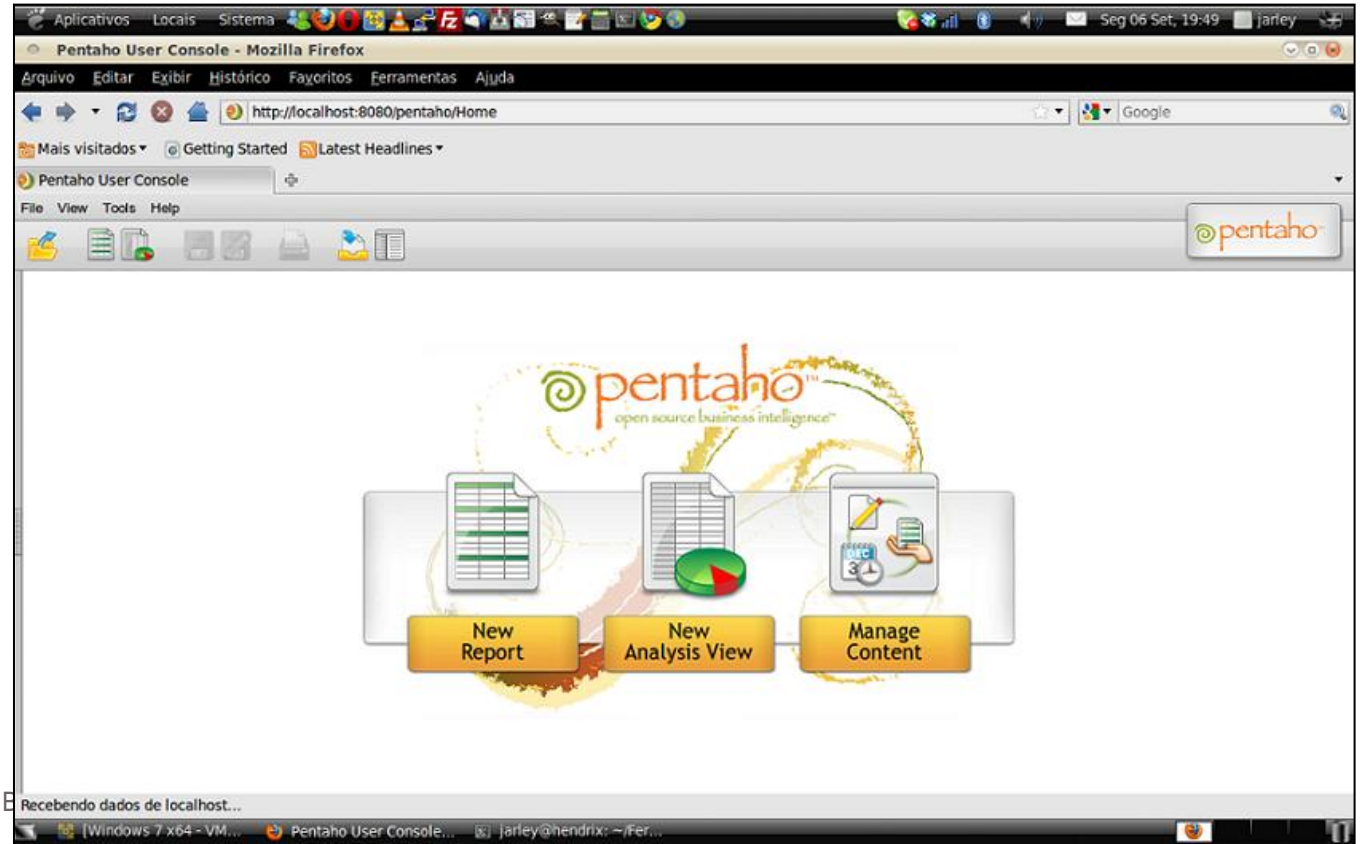
Pentaho BI Suite

- Um pequeno roteiro para rodar o BI Server
  - Acesse a url `http://localhost:8080/pentaho`



Pentaho BI Suite

- Um pequeno roteiro para rodar o BI Server
  - Entre com o usuário "joe" e navegue na aplicação



PENTAHO DATA INTEGRATION

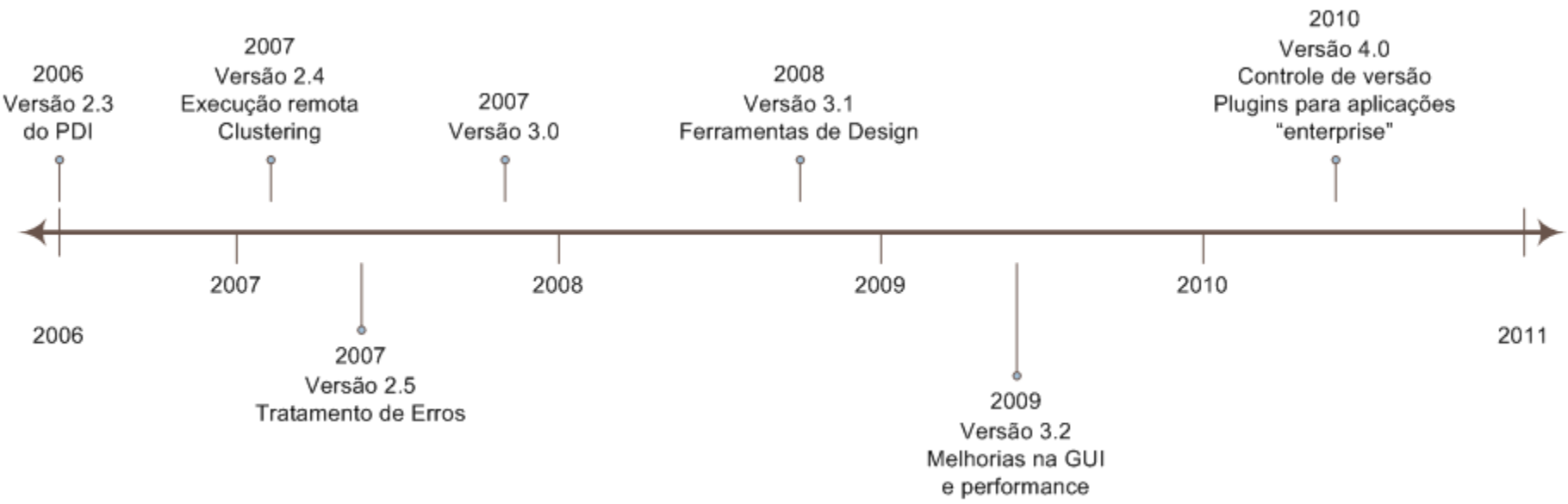
Pentaho Data Integration

Uma das ferramentas de BI da plataforma Pentaho

- Projeto open source encampado pelo Pentaho em 2006
- Desenvolvido por Matt Casters
- Anteriormente conhecido como Kettle
  - **K**DE **E**xtraction, **T**ransportation, **T**ransformation and **L**oading **E**nvironment

Pentaho Data Integration

Timeline do PDI



Pentaho Data Integration

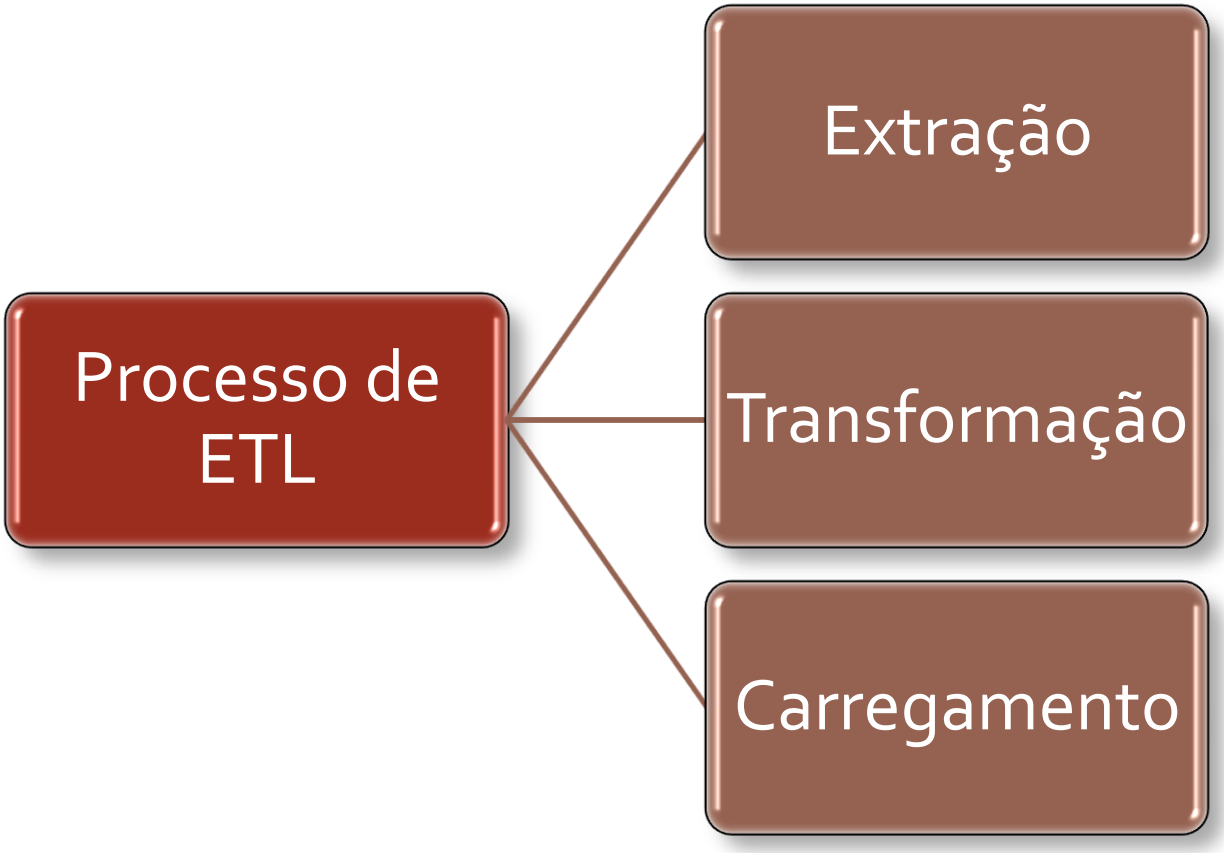
Principais funcionalidades do PDI

Integração de Dados

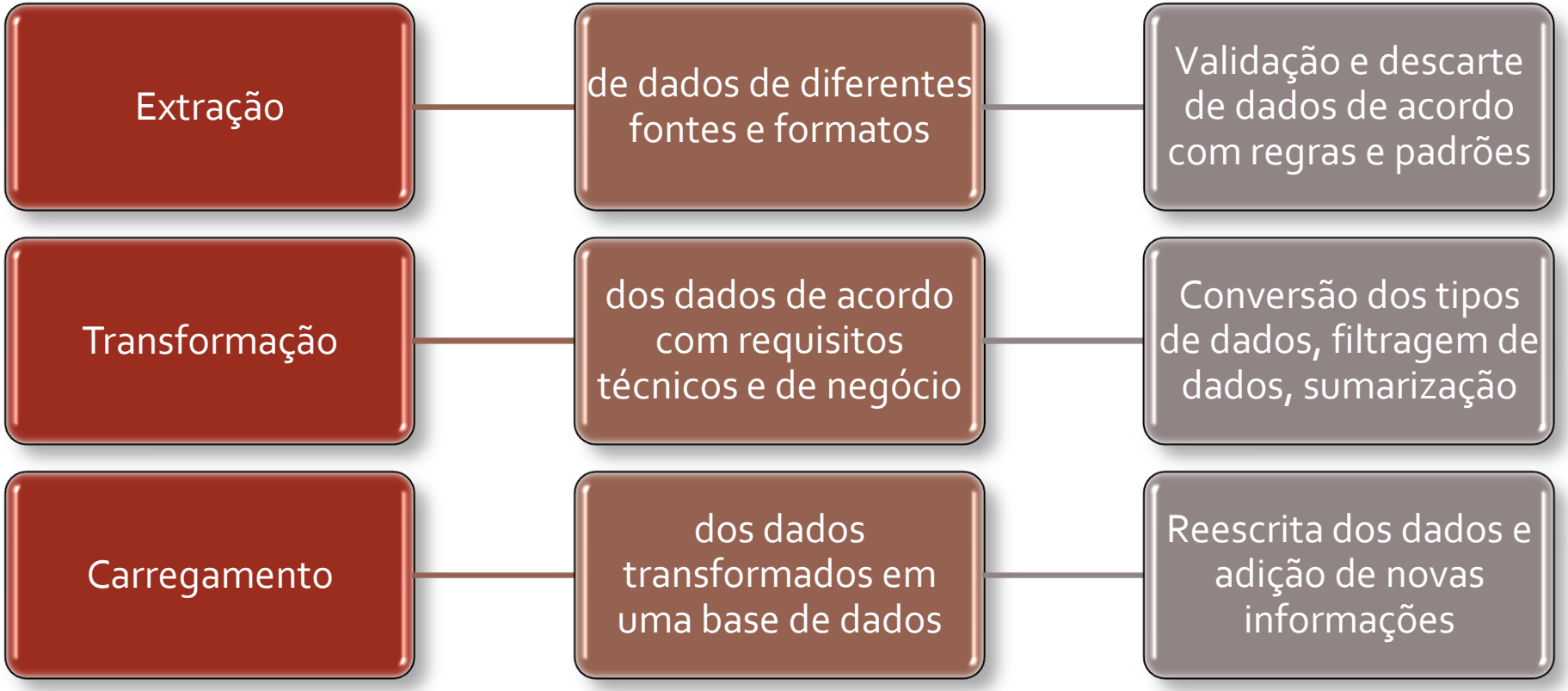
Processo de ETL

Pentaho Data Integration

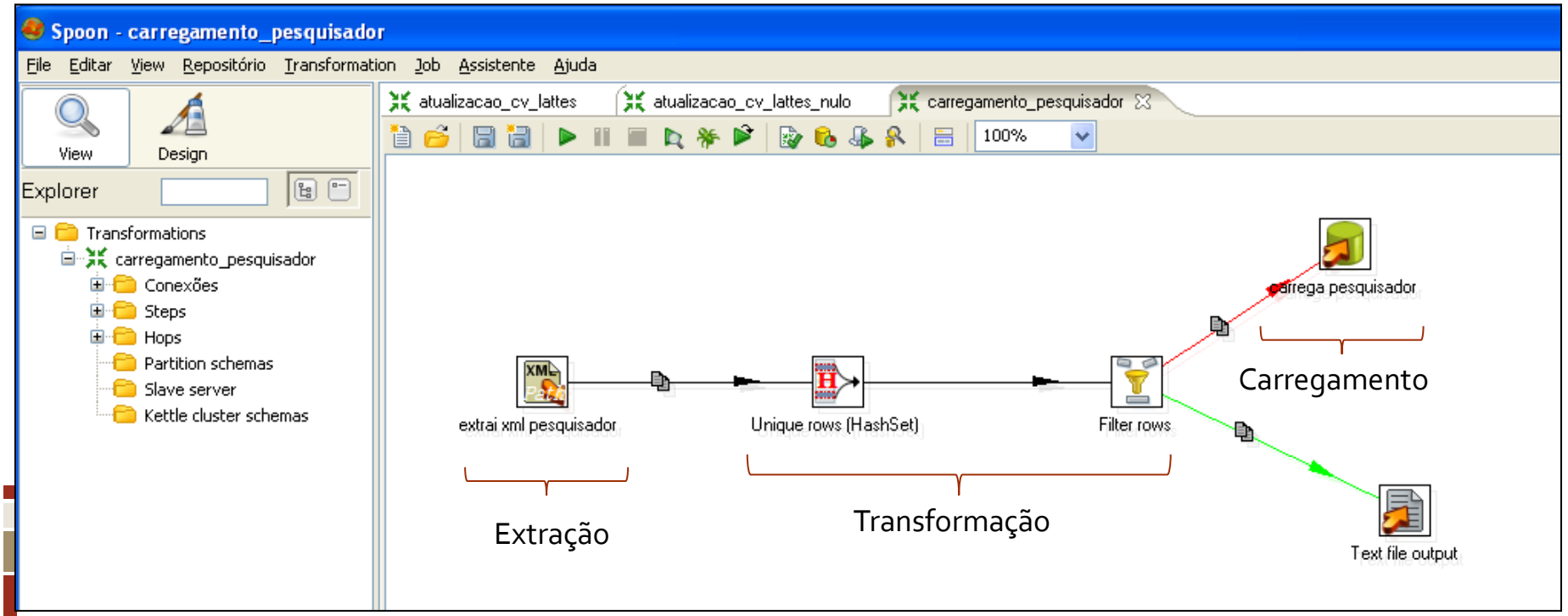
- Carregando dados em um DW ou datamart



Pentaho Data Integration



Pentaho Data Integration



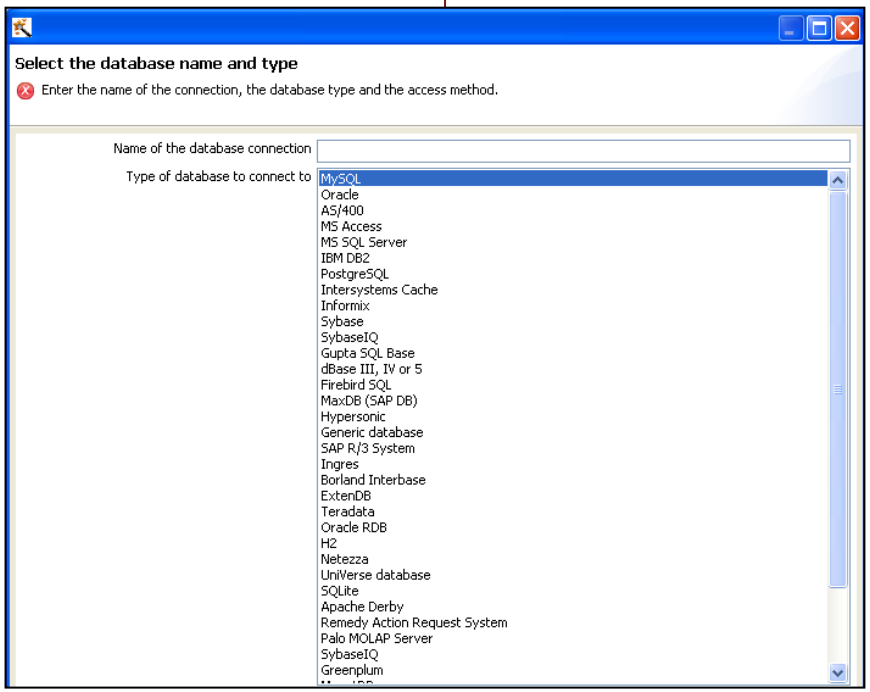
Pentaho Data Integration

- Atividades de Extração
  - Captura dos dados
    - Leitura a partir de diversas fontes
    - Identificação de mudanças desde a última extração.
  - Staging
    - Armazenamento temporário dos dados.

Pentaho Data Integration

Fontes de entrada de dados

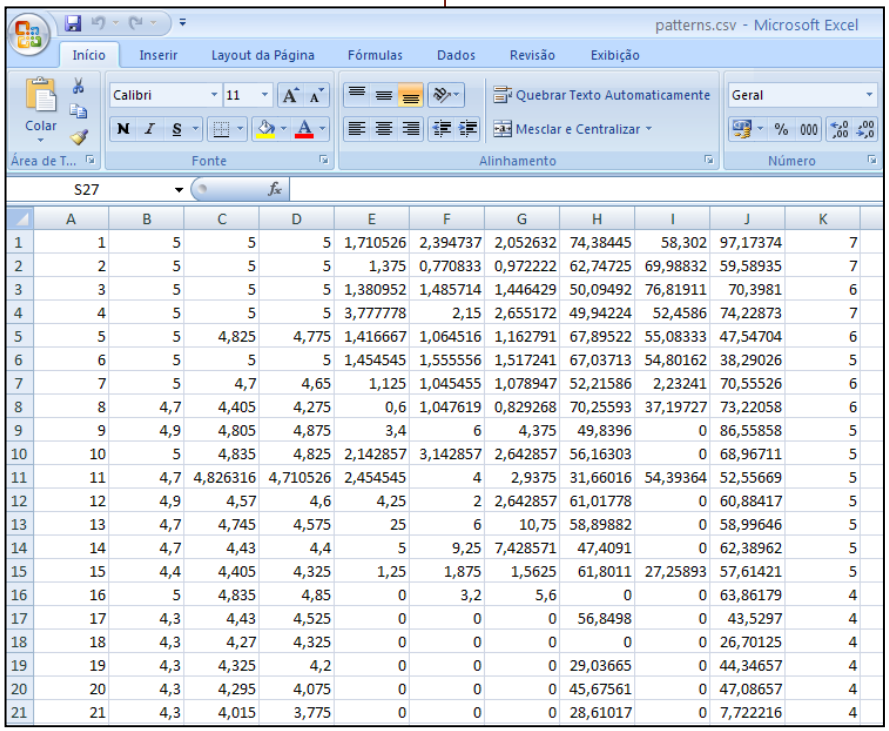
- Sistemas de gerenciamento de banco de dados



Pentaho Data Integration

Fontes de entrada de dados

- Planilhas

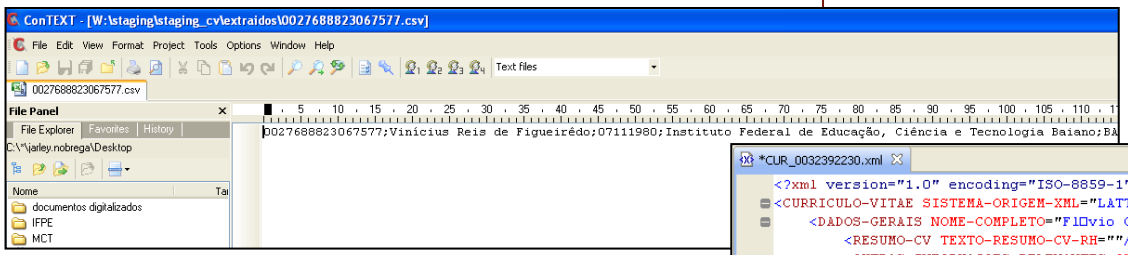


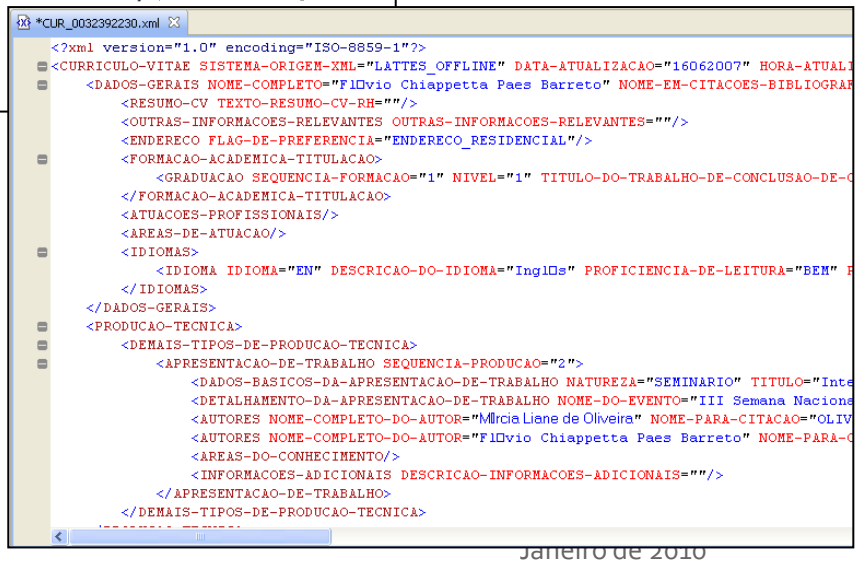
	A	B	C	D	E	F	G	H	I	J	K
1	1	5	5	5	1,710526	2,394737	2,052632	74,38445	58,302	97,17374	7
2	2	5	5	5	1,375	0,770833	0,972222	62,74725	69,98832	59,58935	7
3	3	5	5	5	1,380952	1,485714	1,446429	50,09492	76,81911	70,3981	6
4	4	5	5	5	3,777778	2,15	2,655172	49,94224	52,4586	74,22873	7
5	5	5	4,825	4,775	1,416667	1,064516	1,162791	67,89522	55,08333	47,54704	6
6	6	5	5	5	1,454545	1,555556	1,517241	67,03713	54,80162	38,29026	5
7	7	5	4,7	4,65	1,125	1,045455	1,078947	52,21586	2,23241	70,55526	6
8	8	4,7	4,405	4,275	0,6	1,047619	0,829268	70,25593	37,19727	73,22058	6
9	9	4,9	4,805	4,875	3,4	6	4,375	49,8396	0	86,55858	5
10	10	5	4,835	4,825	2,142857	3,142857	2,642857	56,16303	0	68,96711	5
11	11	4,7	4,826316	4,710526	2,454545	4	2,9375	31,66016	54,39364	52,55669	5
12	12	4,9	4,57	4,6	4,25	2	2,642857	61,01778	0	60,88417	5
13	13	4,7	4,745	4,575	25	6	10,75	58,89882	0	58,99646	5
14	14	4,7	4,43	4,4	5	9,25	7,428571	47,4091	0	62,38962	5
15	15	4,4	4,405	4,325	1,25	1,875	1,5625	61,8011	27,25893	57,61421	5
16	16	5	4,835	4,85	0	3,2	5,6	0	0	63,86179	4
17	17	4,3	4,43	4,525	0	0	0	56,8498	0	43,5297	4
18	18	4,3	4,27	4,325	0	0	0	0	0	26,70125	4
19	19	4,3	4,325	4,2	0	0	0	29,03665	0	44,34657	4
20	20	4,3	4,295	4,075	0	0	0	45,67561	0	47,08657	4
21	21	4,3	4,015	3,775	0	0	0	28,61017	0	7,722216	4

Pentaho Data Integration

Fontes de entrada de dados

- Arquivos texto ou XML





Pentaho Data Integration

- Atividades de Transformação
  - Validação dos dados
    - Verificação se os dados estão corretos e precisos.
    - Filtragem de dados inválidos.
  - Limpeza dos dados
    - Correção de dados inválidos.
  - Decodificação
    - Conversão de atributos (numéricos, categóricos) para adequação a um padrão ou regra.
  - Agregação
  - Geração e gerenciamento de chaves
    - Dimensões identificadas por chaves substitutas (“surrogates”).

Pentaho Data Integration

- Atividades de Carregamento
  - Carregamento das tabelas de fatos
    - Adição de linhas à tabela de fatos.
    - Atualização de atributos de status.
  - Carregamento e manutenção das tabelas de dimensões
    - Adição e atualização de linhas das tabelas de dimensões.

Instalando o PDI

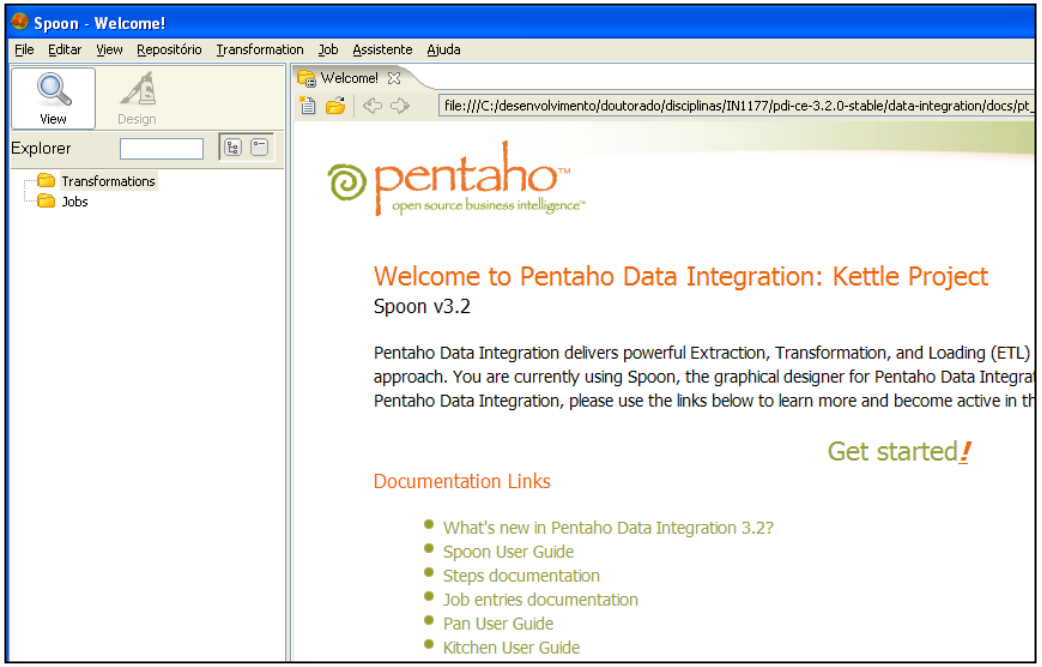
- Pré-requisito
  - JRE (ou JDK) 5.x ou superior.
- Download
  - `http://sourceforge.net/projects/pentaho/files/`
  - Pasta "Data Integration"
  - Obter a última versão estável
    - 4.0.1 – 95.2 MB
    - 3.2.0 – 77.2 MB

Instalando o PDI

- Após descompactar o arquivo
  - Executar spoon.bat ou Kettle.exe (ou spoon.sh no Linux)

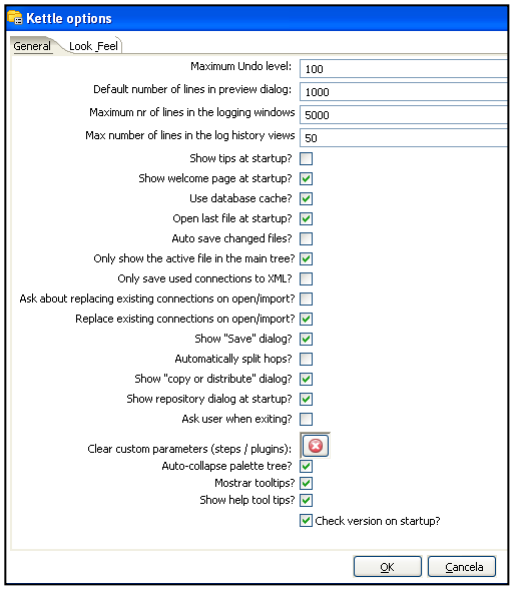
Instalando o PDI

- Clique no botão "No repository"
  - A interface gráfica do PDI (Spoon) será carregada, mostrando uma página de boas vindas.



Instalando o PDI

- Dicas de configuração da área de trabalho do Spoon (Menu Editar -> Opções)
  - Aba "General"
    - Show tips at startup?
    - Show welcome page at startup?
    - ...
  - Aba "Look-and-feel"
    - Preferred language
    - ...



- As mudanças estarão visíveis após reiniciar o Spoon

Principais Componentes do PDI

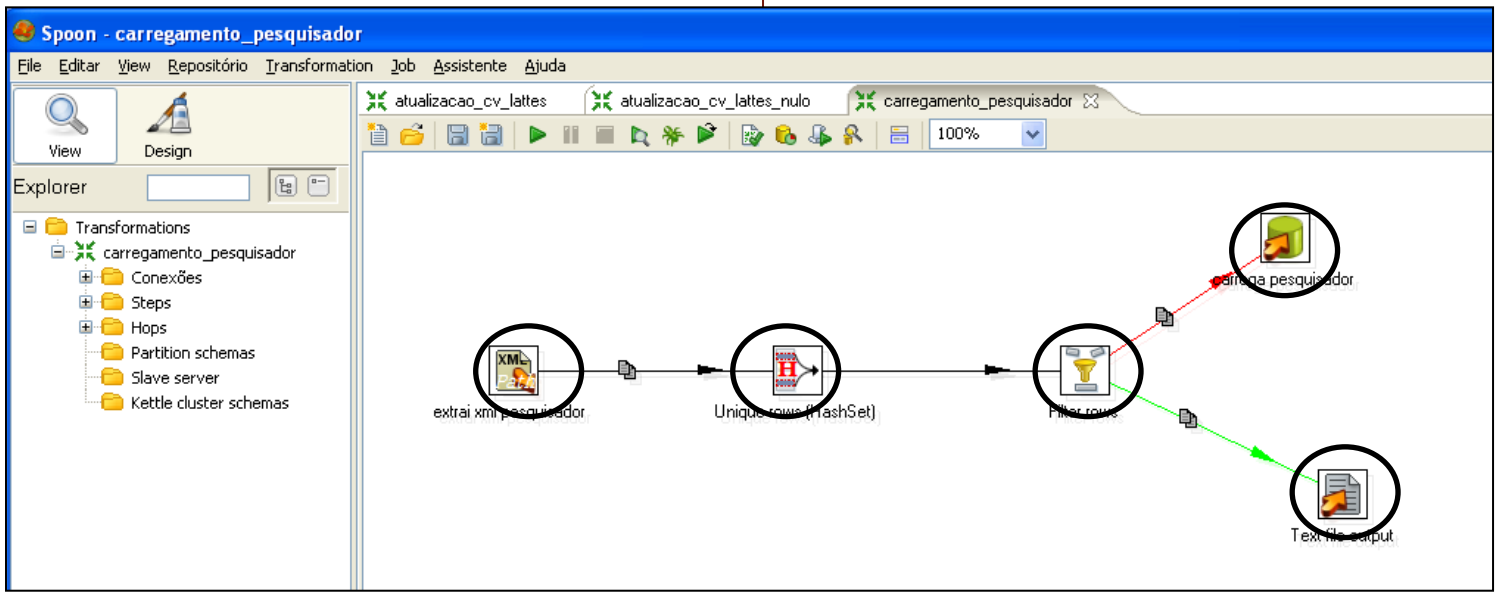
- O PDI trabalha com dois tipos básicos de componentes:
  - Transformações
  - Jobs

- Características de transformações e jobs
  - Definem o fluxo do processo de ETL
  - Contém os metadados do processo de ETL
    - Descrição dos dados;
    - Fontes de entrada e saída;
    - Scheduling;
    - Scripting.

Principais Componentes do PDI

Como as transformações e jobs são executados?

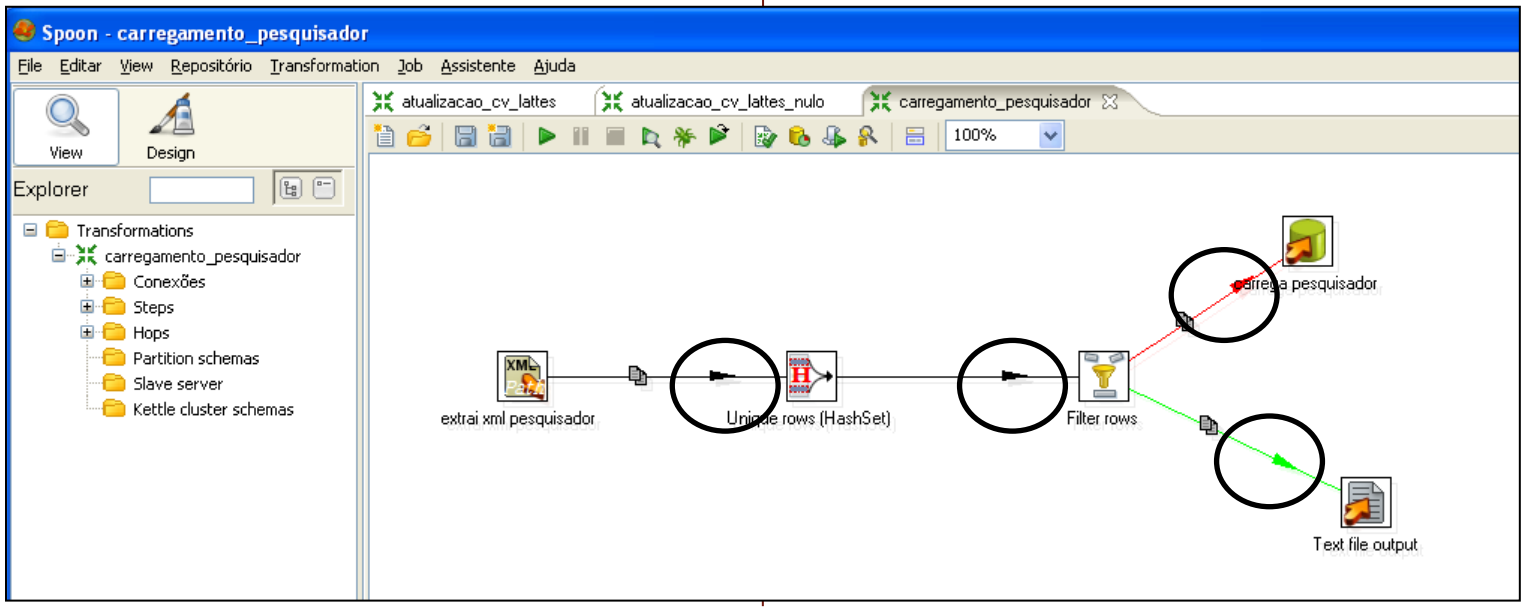
- Uma transformação ou job consiste de uma coleção de itens interconectados



Principais Componentes do PDI

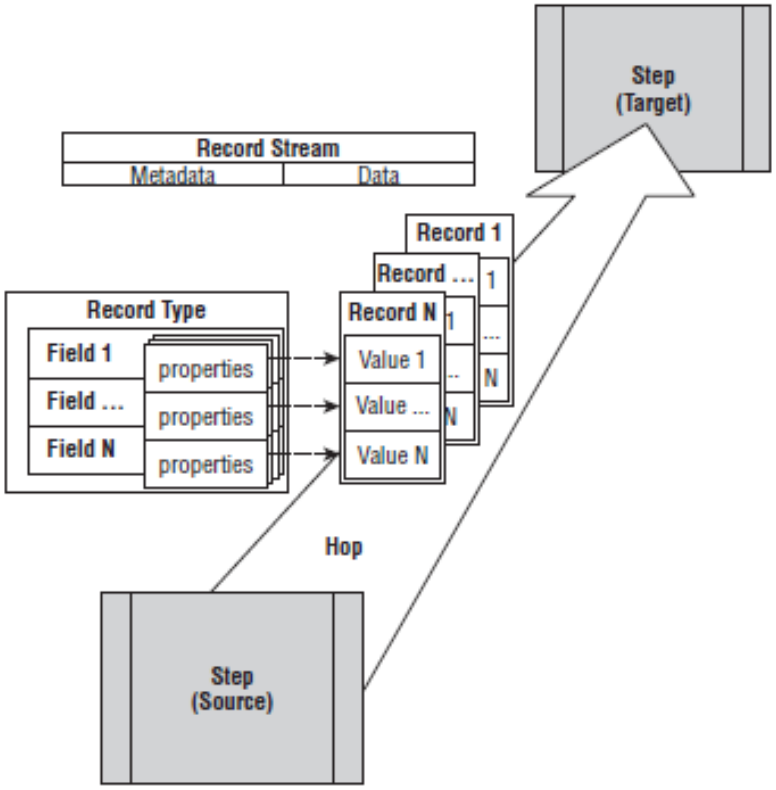
Conexões entre os itens das transformações e jobs

- Hop's
  - Pipeline do fluxo de registros



Principais Componentes do PDI

- Steps, hops e o fluxo de registros



(Bouman and Dongen, 2009)

Principais Componetes do PDI

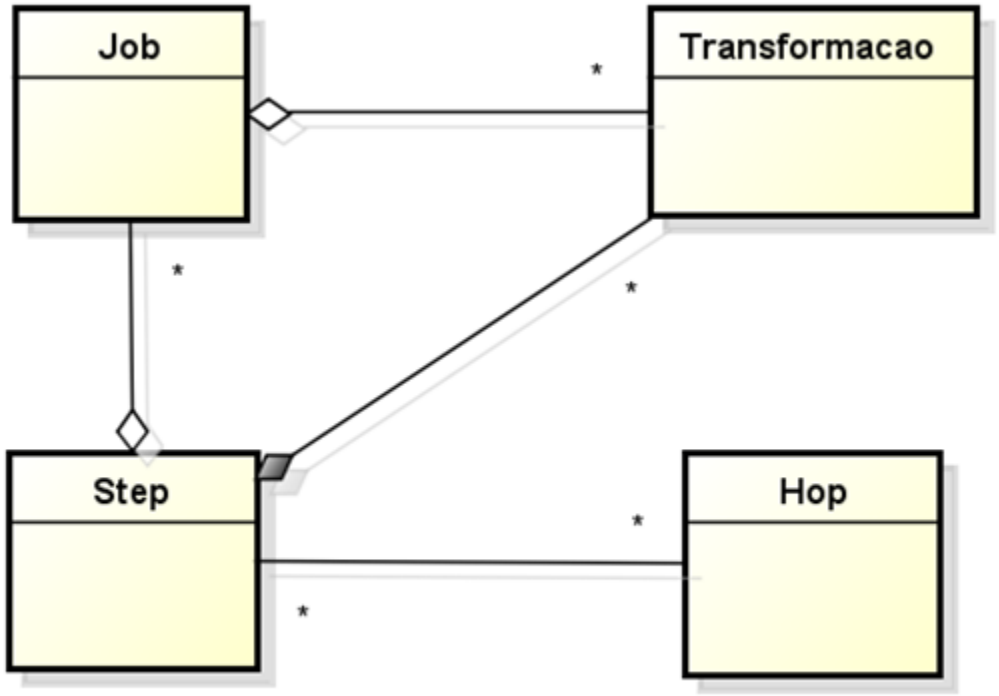
Transformações

- Consiste de uma coleção de *steps de transformação*
- Cada *step* denota uma operação do processo de ETL
- A saída de um *step* produz um conjunto de registros
- Fluxo dos *steps* da transformação ocorre de forma simultânea e assíncrona
- Arquivo `.ktr`

Jobs

- Consiste de uma coleção de transformações ou de *steps de jobs*
- Cada *entrada do job* denota uma tarefa do processo de ETL
- A saída de cada entrada do job produz um status de execução
- Fluxo dos *steps* do job ocorre de forma sequencial
- Arquivo `.kjb`

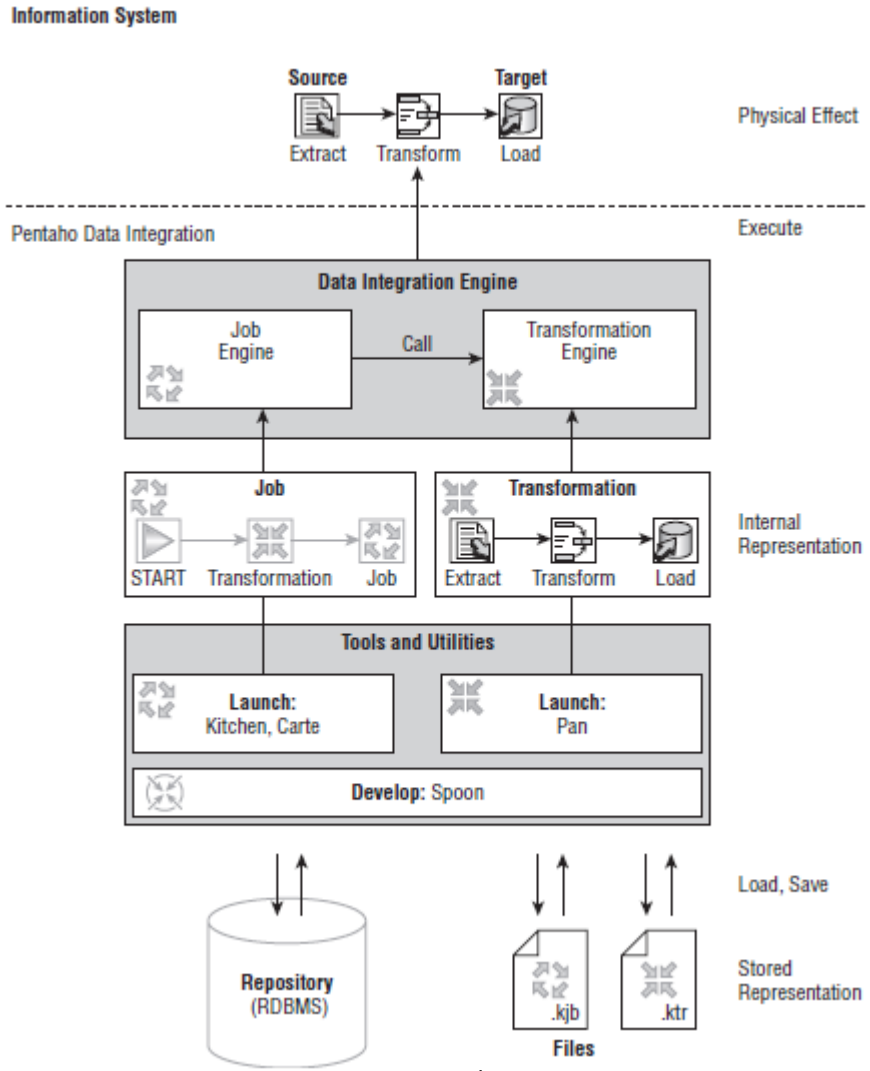
Metamodelo dos componentes do PDI



Principais Componentes do PDI

- Outros componentes do PDI:
  - Repositórios
    - Os metadados das transformações e jobs podem ser persistidos em um banco de dados (repositório)
  - Ferramentas:
    - Spoon: IDE para desenvolvimento visual.
    - Pan: execução de transformações em linha de comando.
    - Kitchen: execução de jobs em linha de comando.
    - Carte: servidor de para execução remota de transformações e jobs.

Arquitetura do PDI



(Bouman and Dongen, 2009)

Exercícios 1 e 2

- Criando as primeiras transformações no PDI
  - Transformação simples
  - Processo de ETL
    - Extração de dados de uma fonte (arquivo texto)
    - Transformação dos dados
    - Carregamento dos dados transformados (arquivo texto)

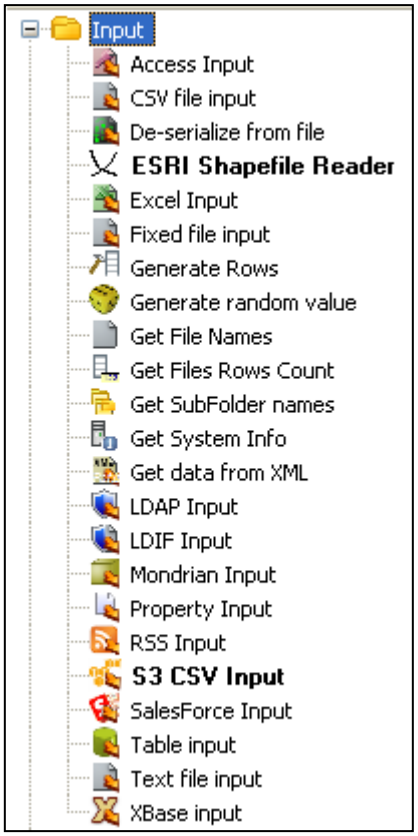
Exercício 3

- Criando uma conexão com um banco de dados

TRABALHANDO COM ARQUIVOS NO PDI

Extraindo dados no PDI

- Vários steps para extrair dados
  - Banco de dados;
  - Informações do sistema;
  - Arquivos texto;
  - ...

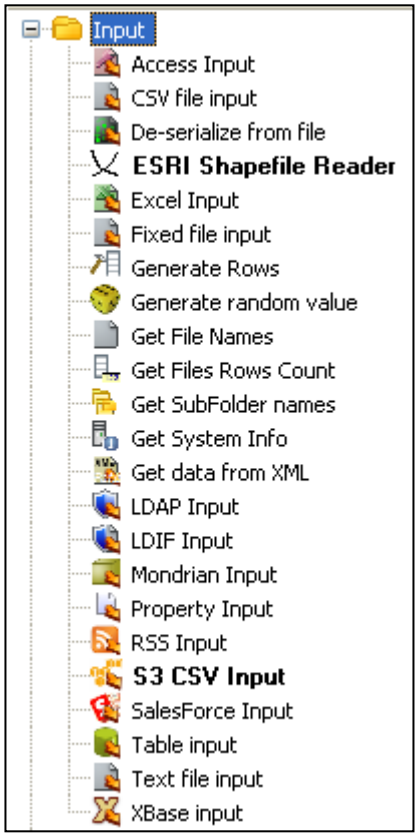


Extraindo dados no PDI

- Principais propriedades dos steps de extração
  - Nome do step
    - Obrigatório e único para cada step
  - Nome e localização do arquivo
  - Descrição do conteúdo
    - Separador, codificação, cabeçalho, etc.
    - Depende do tipo do step
  - Campos
  - Filtros
    - Ex.: pular linhas em branco, ler as primeiras *n* linhas, etc.

Extraindo dados de arquivos

- Modo "primitivo" de armazenar dados
  - Comma-separated values (CSV);
  - Planilhas;
  - Arquivos flat;
  - XML.



Extraindo dados de arquivos

Nome do Step	Fonte dos dados
CSV File Input	Campos de um arquivo .CSV
Excel Input	Células de uma planilha .XLS
Fixed file input	Texto de tamanho fixo
Text file input	Idem ao CSV + tratamento de erros + filtros
Get data from XML	Nós e atributos de tags no formato XML

Exercício 4

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

Lendo vários arquivos

- Até agora extraímos dados de um único arquivo texto
  - Extração de dados de vários arquivos:
    - Lista de arquivos
    - Expressões regulares

Exercícios 5 e 6

- Adicionando uma lista de arquivos de entrada.
- Usando expressões regulares

Expressões regulares

- Em vários steps do PDI podemos usar expressões regulares
- Exemplos

Expressão regular	Combina com...	Exemplos
.*\.txt	Qualquer arquivo .txt	Arquivo.txt
test(19|20)\d\d-(0[1-9]|1[012])\.txt	Qualquer arquivo começando com `test`, seguido por uma data usando o formato yyyy-mm	test2009-12.txt test2009-01.txt
(?i)test.+\.txt	Qualquer arquivo .txt começando com `test` escrito em maiúsculo ou minúsculo	TeSTcaseinsensitive.tXt

Expressões regulares

- Para saber mais sobre expressões regulares
  - Regular Expression Quick Start: http://www.regular-expressions.info/quickstart.html
  - The Java Regular Expression Tutorial: http://java.sun.com/docs/books/tutorial/essential/regex/
  - Java Regular Expression Pattern Syntax: http://java.sun.com/javase/6/docs/api/java/util/regex/Pattern.html

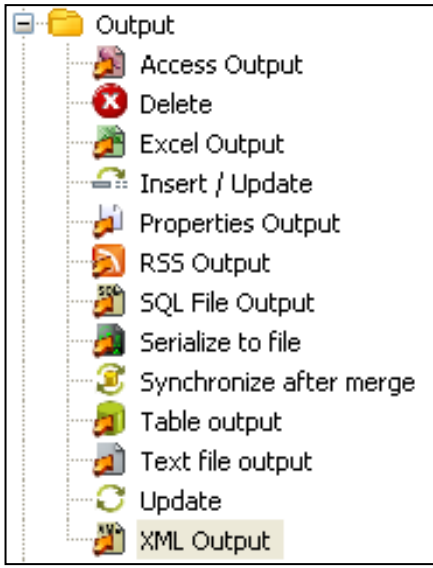
Enviando dados para arquivos

- Vários steps para enviar dados para arquivos

Nome do Step	Destino dos dados
Excel output	Células de uma planilha no formato .xls
SQL file output	Comandos SQL em arquivo texto
Text file output	Linhas em um arquivo texto (txt ou CSV)
XML output	Nós e atributos de tags no formato XML

Enviando dados para arquivos

- Principais propriedades
  - Nome do step
    - Obrigatório e único para cada step
  - Nome e localização do arquivo
    - Opção *Append*
  - Descrição do conteúdo
    - Separador, codificação, cabeçalho, etc.
    - Depende do tipo do step
  - Campos



Definições de dados do PDI

Dois conceitos importantes de dados para o PDI

Rowset

Streams

Definições de dados do PDI

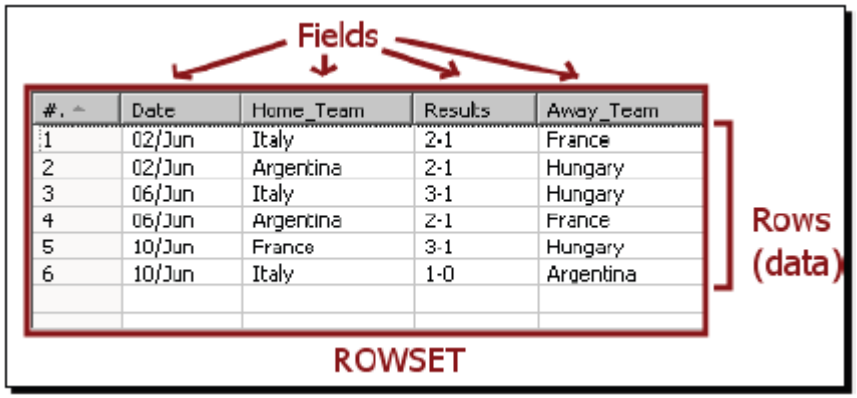
Rowset

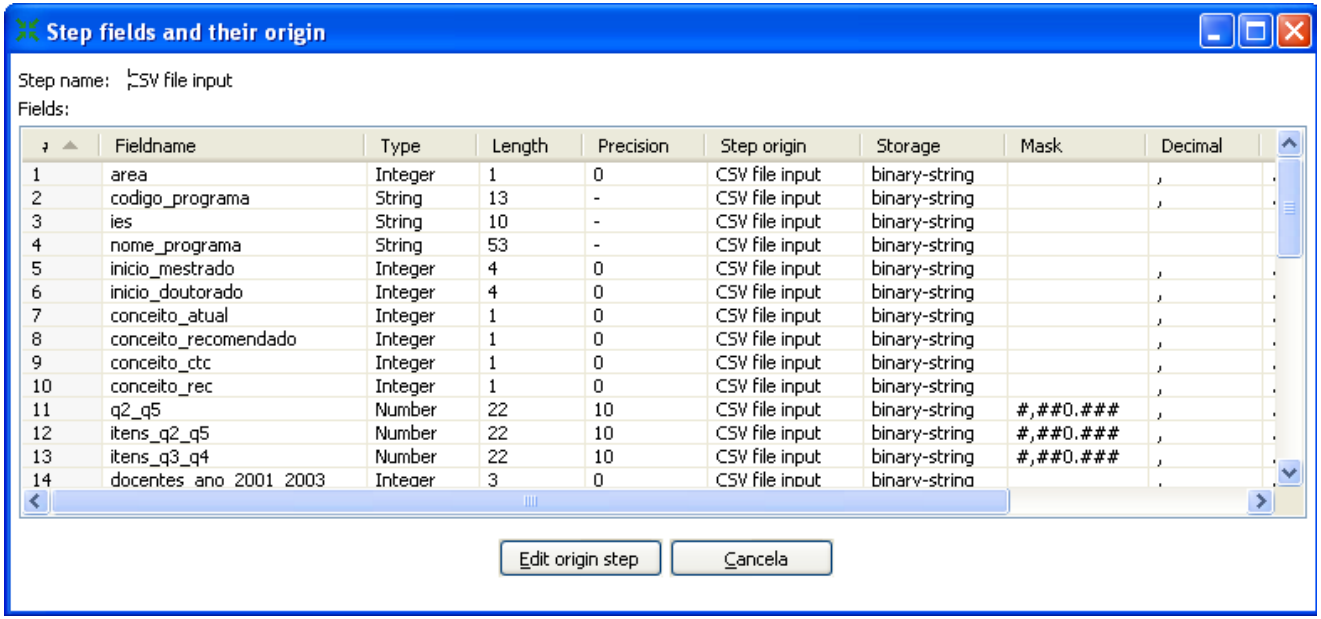
- Dados representados de forma tabular (datasets)
- Cada coluna representa um campo
  - Nome (obrigatório)
  - Tipo: `Number` (float), `String`, `Date`, `Boolean`, `Integer` e `Big Number`
- Cada linha corresponde a um membro do dataset

Streams

- Dados enviados de um step para outro
  - Os *hops* apenas repassam o fluxo de dados
- Cada step pode ter um *rowset* de entrada e outro de saída
- Botão direito -> Mostra campos de entrada/saída

Definições de dados do PDI

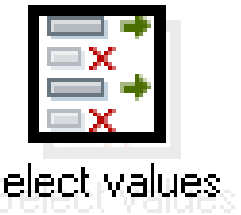




Transformações no dataset de arquivos

- A forma mais simples de fazer transformações no rowset de um arquivo
  - Step Select Values

- Operações básicas
  - Selecionar e Alterar Campos
  - Remover Campos
  - Alterar metadados dos campos



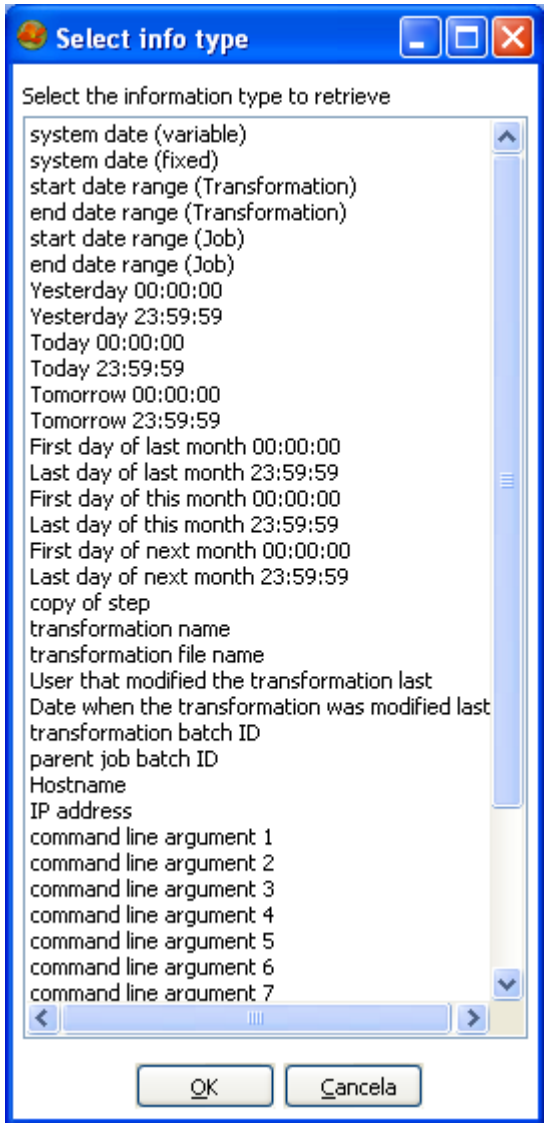
Exercício 7

- Alterando os campos do Exercício 6
- Gerando a saída para uma planilha Excel

Extraindo informações do Ambiente

- O que vimos até agora?
  - Extração dos dados de arquivos
  - Realizando transformações básicas
  - Carregando dados em arquivos
- Como obter dados sem conhecer previamente o nome dos arquivos?
  - Lendo informações do ambiente
  - Step *Get System Info*

Extraindo informações do Ambiente



- Variáveis do S.O.
  - Datas
  - Sistema de arquivos
  - Rede
- Variáveis de ambiente do PDI

Exercício 8

- Extraindo informações do sistema

Tipos de Dados

- Todo campo de um dataset possui um tipo de dado
- Tipos suportados pelo PDI
  - Number (float)
  - String
  - Date
  - Boolean
  - Integer
  - Big Number

Tipos de Dados

- Date (padrão API Java)

Letter	Date or Time Component	Presentation	Examples
G	Era designator	Text	AD
y	Year	Year	1996; 96
M	Month in year	Month	July; Jul; 07
w	Week in year	Number	27
W	Week in month	Number	2
D	Day in year	Number	189
d	Day in month	Number	10
F	Day of week in month	Number	2
E	Day in week	Text	Tuesday; Tue
a	Am/pm marker	Text	PM
H	Hour in day (0-23)	Number	0
k	Hour in day (1-24)	Number	24
K	Hour in am/pm (0-11)	Number	0
h	Hour in am/pm (1-12)	Number	12
m	Minute in hour	Number	30
s	Second in minute	Number	55
S	Millisecond	Number	978
z	Time zone	General time zone	Pacific Standard Time; PST; GMT-08:00
Z	Time zone	RFC 822 time zone	-0800

Tipos de Dados

- Date - Exemplos

Date and Time Pattern	Result
"yyyy.MM.dd G 'at' HH:mm:ss z"	2001.07.04 AD at 12:08:56 PDT
"EEE, MMM d, ''yy"	Wed, Jul 4, '01
"h:mm a"	12:08 PM
"hh 'o''clock' a, zzzz"	12 o'clock PM, Pacific Daylight Time
"K:mm a, z"	0:08 PM, PDT
"yyyyy.MMMMM.dd GGG hh:mm aaa"	02001.July.04 AD 12:08 PM
"EEE, d MMM yyyy HH:mm:ss Z"	Wed, 4 Jul 2001 12:08:56 -0700
"yyMMddHHmmssZ"	010704120856-0700
"yyyy-MM-dd'T'HH:mm:ss.SSSZ"	2001-07-04T12:08:56.235-0700

Formato padrão: `yyyy/MM/dd`

Tipos de Dados

- Campos numéricos (padrão API Java)
  - O PDI tenta "interpretar" dados numéricos
  - Campos mais elaborados precisam de um formato
    - Formatos mais usados

Símbolo	Significado
#	Dígito zero não é mostrado (pode arredondar)
0	Se o dígito não estiver presente, o zero é mostrado no lugar
.	Separador decimal
–	Sinal de menos
%	Campo deve ser multiplicado por 100 e exibido como percentual

Tipos de Dados

- Campos numéricos (padrão API Java)
  - Exemplos - campo com valor `99.55`

Formato	Resultado
`#`	`100` (arredondamento)
`0`	`100` (arredondamento)
`#.#`	`99.6`
`#.##`	`99.55`
`#.000`	`99.550`
`000.000`	`099.550`

Tipos de Dados

- Campos numéricos (padrão API Java)
  - Algumas considerações:
    - Se não especificar o formato -> informar tamanho e precisão
    - Por padrão, o PDI tenta "interpretar" o número e repassa pelo *hop* sem aplicar nenhum formato.

Exercício 9

- Aplicando formatos para datas e números do Exercício 8

Arquivos XML

- Arquivos (ou documentos) XML são utilizados para:
  - Armazenar dados
  - Troca de dados entre sistemas heterogêneos

- Entrada de dados XML
  - Step *Get data from XML*

- Saída de dados XML
  - Step *XML output*

Arquivos XML

- Como o PDI trata arquivos XML?

```
<?xml version="1.0" encoding="UTF-8"?>
<world>
...
  <country>
    <name>Argentina</name>
    <capital>Buenos Aires</capital>
    <language isofficial="T">
      <name>Spanish</name>
      <percentage>96.8</percentage>
    </language>
    <language isofficial="F">
      <name>Italian</name>
      <percentage>1.7</percentage>
    </language>
    <language isofficial="F">
      <name>Indian Languages</name>
      <percentage>0.3</percentage>
    </language>
  </country>
...
</world>
```

elemento

atributo

Arquivos XML

- Como o PDI trata arquivos XML?
  - Step *Get data from XML*
    - *Notação Xpath:* Conjunto de regras para recuperar informação de um documento XML
    - Documento XML tratado como uma árvore formada por nós.
    - Tipos de nós:
      - Elementos;
      - Atributos;
      - Texto

Arquivos XML

- Como o PDI trata arquivos XML?
  - Relacionamento entre os nós
    - Um nó tem um pai
    - Um nó tem zero ou mais filhos, irmãos, ancestrais ou descendentes

Arquivo de exemplo: `country` é o pai dos elementos `name`, `capital` e `language`. Os três elementos são filhos de `country`.

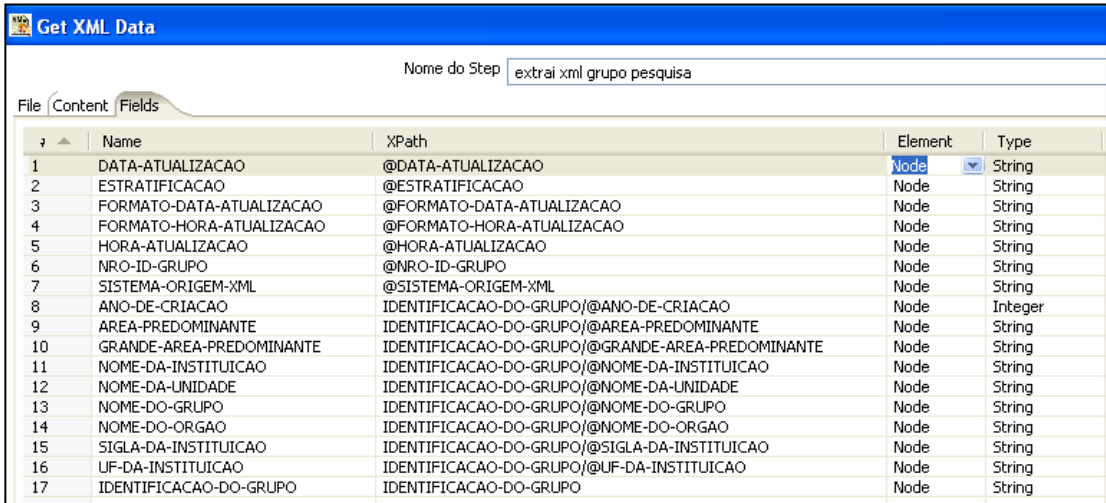
Arquivos XML

- Como o PDI trata arquivos XML?
  - Para acessar um nó
    - Usar uma expressão no formato XPath relativa ao nó corrente.

Arquivos XML

- Exemplos XPath

Expressão	Descrição
`node_name`	Seleciona todos os nós filhos do nó `node_name`.
`.`	Seleciona o nó corrente
`..`	Seleciona o pai do nó corrente
`@`	Seleciona um atributo



#	Name	XPath	Element	Type
1	DATA-ATUALIZACAO	@DATA-ATUALIZACAO	Node	String
2	ESTRATIFICACAO	@ESTRATIFICACAO	Node	String
3	FORMATO-DATA-ATUALIZACAO	@FORMATO-DATA-ATUALIZACAO	Node	String
4	FORMATO-HORA-ATUALIZACAO	@FORMATO-HORA-ATUALIZACAO	Node	String
5	HORA-ATUALIZACAO	@HORA-ATUALIZACAO	Node	String
6	NRO-ID-GRUPO	@NRO-ID-GRUPO	Node	String
7	SISTEMA-ORIGEM-XML	@SISTEMA-ORIGEM-XML	Node	String
8	ANO-DE-CRIACAO	IDENTIFICACAO-DO-GRUPO/@ANO-DE-CRIACAO	Node	Integer
9	AREA-PREDOMINANTE	IDENTIFICACAO-DO-GRUPO/@AREA-PREDOMINANTE	Node	String
10	GRANDE-AREA-PREDOMINANTE	IDENTIFICACAO-DO-GRUPO/@GRANDE-AREA-PREDOMINANTE	Node	String
11	NOME-DA-INSTITUICAO	IDENTIFICACAO-DO-GRUPO/@NOME-DA-INSTITUICAO	Node	String
12	NOME-DA-UNIDADE	IDENTIFICACAO-DO-GRUPO/@NOME-DA-UNIDADE	Node	String
13	NOME-DO-GRUPO	IDENTIFICACAO-DO-GRUPO/@NOME-DO-GRUPO	Node	String
14	NOME-DO-ORGAO	IDENTIFICACAO-DO-GRUPO/@NOME-DO-ORGAO	Node	String
15	SIGLA-DA-INSTITUICAO	IDENTIFICACAO-DO-GRUPO/@SIGLA-DA-INSTITUICAO	Node	String
16	UF-DA-INSTITUICAO	IDENTIFICACAO-DO-GRUPO/@UF-DA-INSTITUICAO	Node	String
17	IDENTIFICACAO-DO-GRUPO	IDENTIFICACAO-DO-GRUPO	Node	String

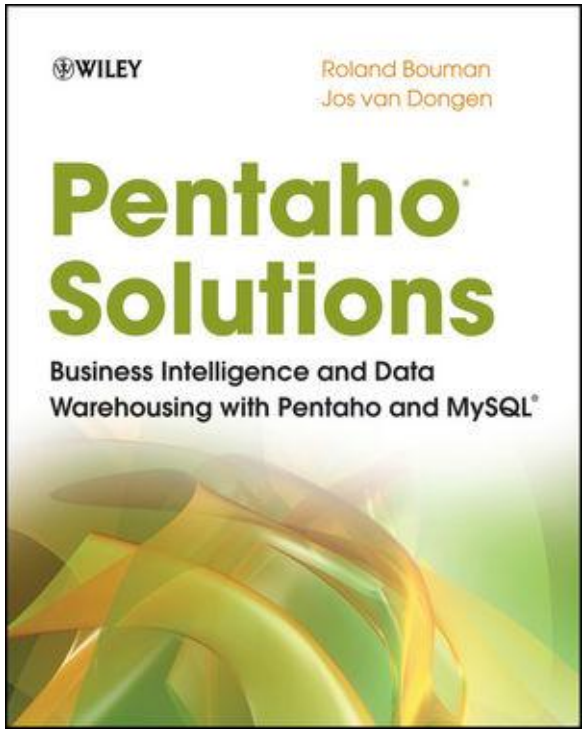
Exercício 10

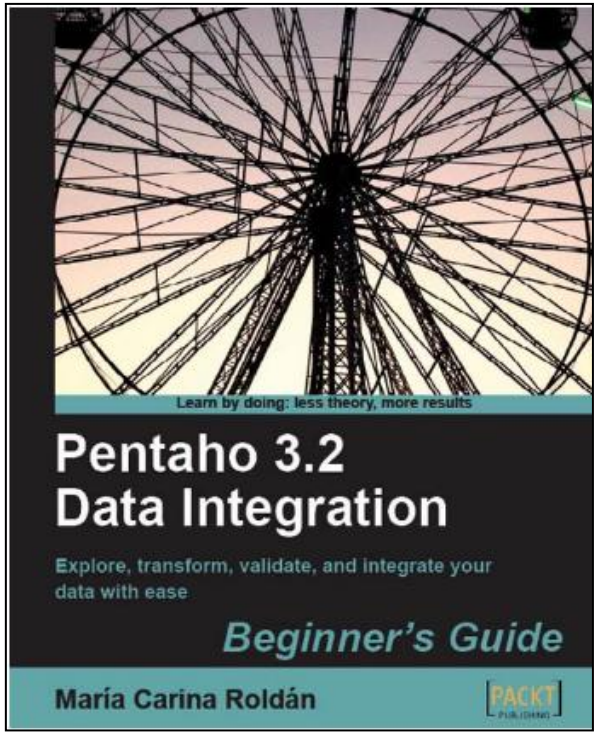
- Extraindo uma lista com dados de países em XML
- Salvando o resultado em uma planilha

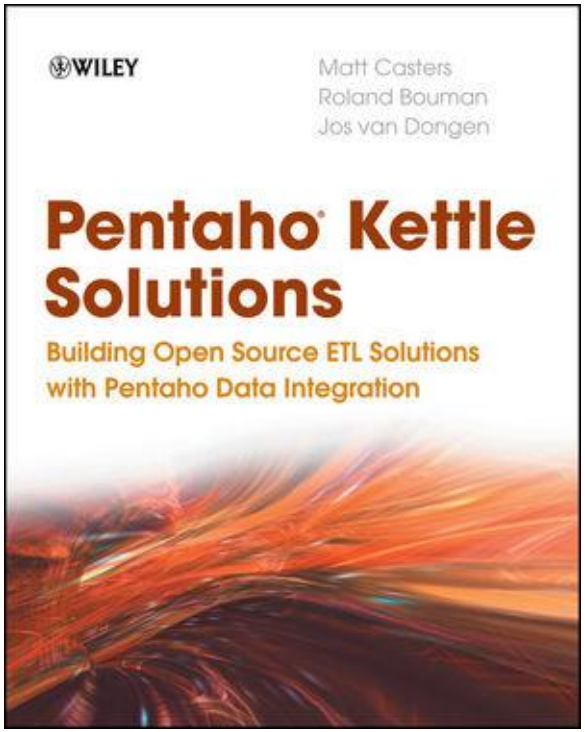
Resumo da Semana 1

- Arquitetura do Pentaho BI server
- Instalação do PDI
- Arquitetura do PDI
- Extração de dados em arquivos texto (plain e XML)
- Carregamento de dados em arquivos texto e planilhas
- Extração de informação a partir de informações do ambiente
- Tipos de dados suportados pelo PDI
- Operações básicas de transformações

Bibliografia







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