

Pentaho Data Integration

PDI

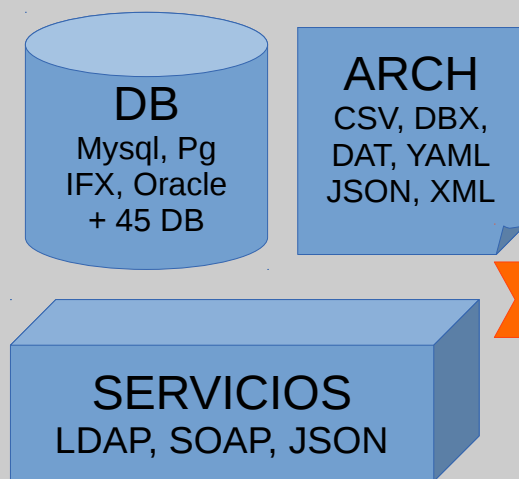
Herramienta de la solución PENTAHO
que se ocupa de los procesos ETL

Se utiliza para poblar DWH
Migrar Datos
Exportar vistas

ETL

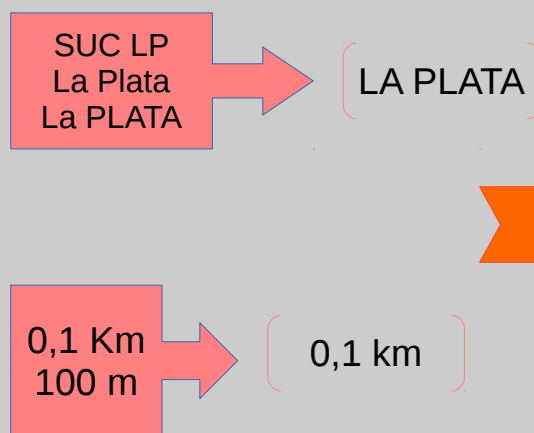
E Extraction

Extrae datos de una o mas fuentes de datos de origen



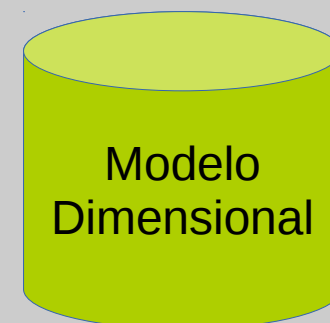
T Transformation

Transforma los datos para que sean consistentes y fáciles de interpretar



L Load

Carga los datos en un almacén unificado y con una estructura que simplifica el acceso



PDI

JAVA - Multiplataforma

PASOS

EJECUTAR
DETENER

AREA
DE
TRABAJO

INFO
EJECUCION

Spoon - [EE Repository] Sample Transformation v1.20

File Edit View Action Tools Help

View Design

Steps

Input

- Access Input
- CSV file input
- Data Grid
- De-serialize from file
- ESRI Shapefile Reader
- Excel Input
- Fixed file input
- Generate random value
- Generate Rows
- Get data from XML
- Get File Names
- Get Files Rows Count
- Get SubFolder names
- Get System Info
- Google Analytics Input
- Google Docs Input
- LDAP Input
- LDIF Input
- Mondrian Input
- OLAP Input
- Property Input
- RSS Input
- S3 CSV Input
- Salesforce Input
- SAP Input
- Table input
- Text file input
- XBase input

Output

- Transform
- Utility
- Flow
- Scripting
- Lookup
- Joins
- Data Warehouse
- Validation
- Statistics

Sample Transformation

Read Sales Data

Filter Missing Zips

Value Mapper

Select values

Number range

Write to Database

Read Postal Codes

Lookup Missing Zips

Prepare Field Layout

To test this transformation, you will need to:

- Make sure the Hypersonic sample database is running
(./pdi-ee\data-integration-server\data\start_hypersonic.bat)
- Open the Table Output step and click the SQL button to create the target output table

Execution Results

Execution History Logging Step Metrics Performance Graph

Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	inpu
1 Filter Missing Zips	0	2823	2823	0	0	0	0	0	Finished	0.5	6019.1	
2 Lookup Missing Zips	0	21455	76	0	0	0	0	0	Finished	0.9	24520.0	
3 Read Postal Codes	0	0	21379	21380	0	1	0	0	Finished	0.7	31815.4	
4 Prepare Field Layout	0	76	76	0	0	0	0	0	Finished	0.9	85.2	
5 Value Mapper	0	2823	2823	0	0	0	0	0	Finished	0.9	3112.4	
6 Read Sales Data	0	0	2823	2824	0	1	0	0	Finished	0.3	8209.3	
7 Select values	0	2823	2823	0	0	0	0	0	Finished	0.9	3112.4	
8 Number range	0	2823	2823	0	0	0	0	0	Finished	0.9		
9 Write to Database	0	2823	2823	0	2823	0	0	0	Finished	1.1		

PDI

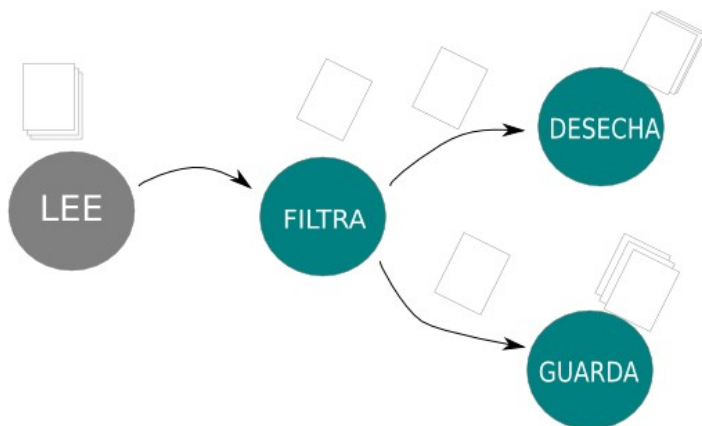
Transformaciones

Kettle Transformation (*.ktr)

Cada paso se ejecuta concurrentemente.

El flujo es a nivel de registro

Cada registro, avanza en el flujo independientemente de los demás



Trabajos

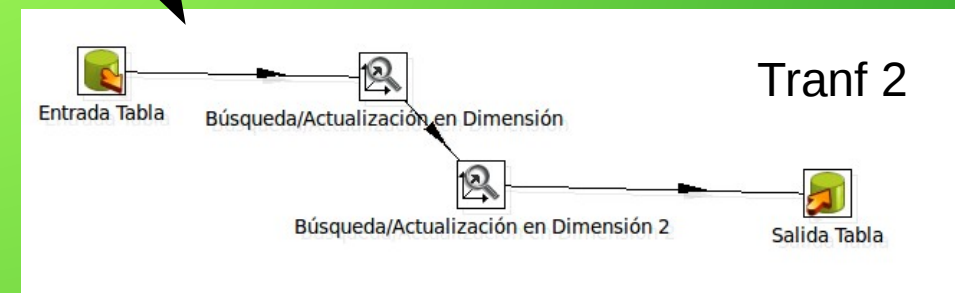
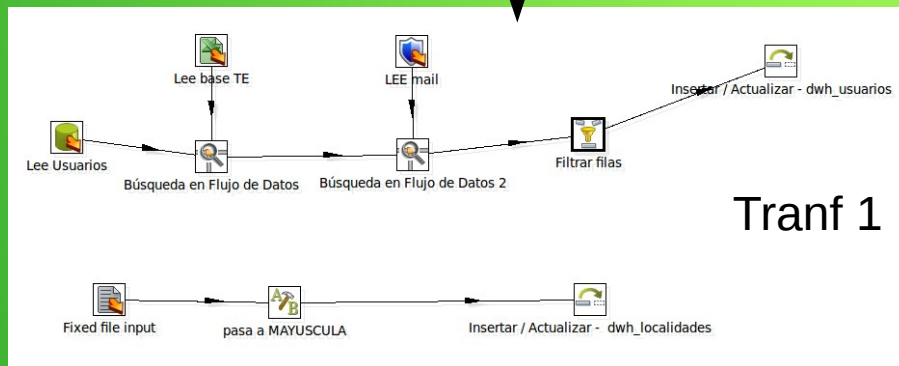
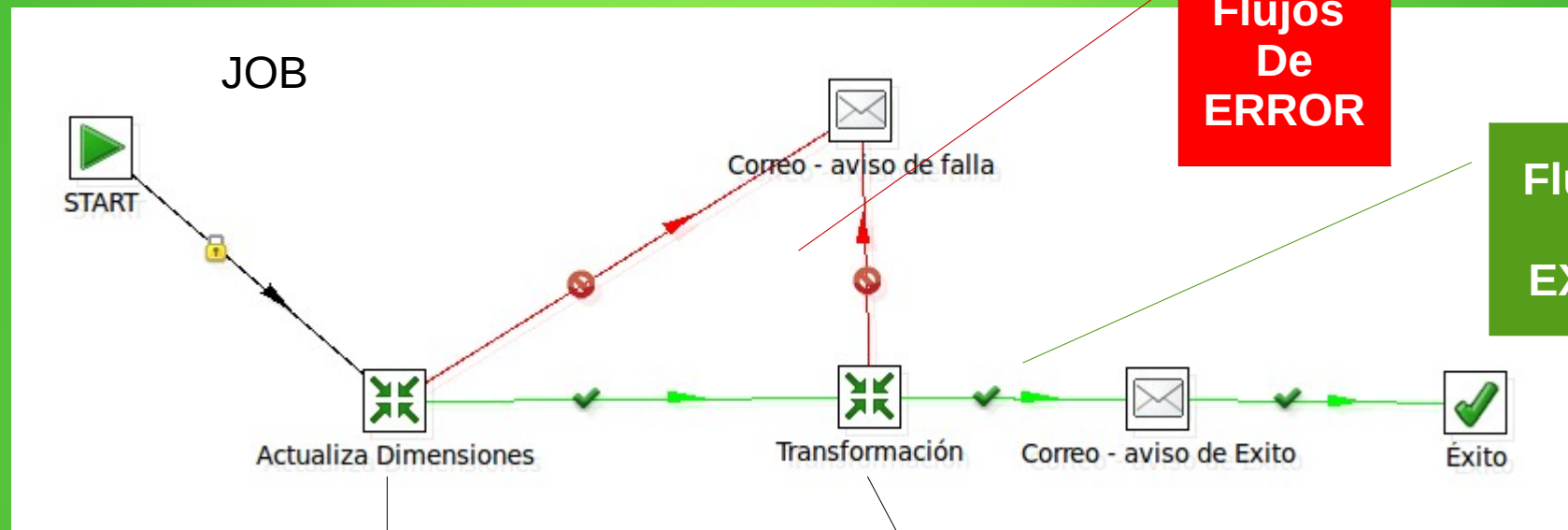
Kettle Job *.kjb

Se ejecuta de un paso a la vez.

Se utiliza para organizar varias transformaciones



PDI

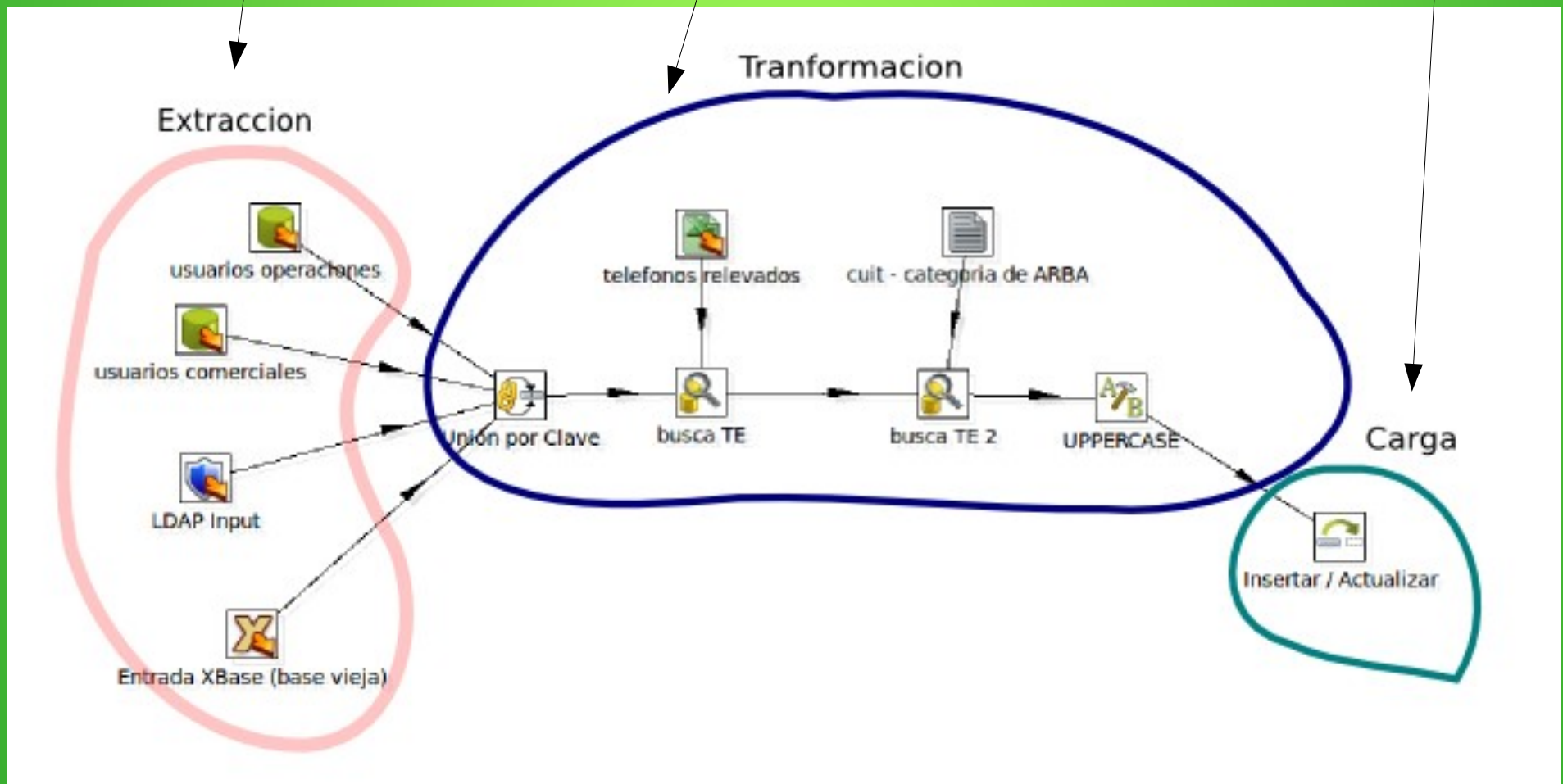


PDI

Pasos de ENTRADA
Generan filas

Pasos de TRANSF
Modifican filas

Pasos de SALIDA
GUARDAN filas



PDI Pasos de ENTRADA

Consulta a Base de Datos



Tabla - Entrada

Nombre paso: Entrada Tabla

Conexión: guarani_ol_agro_sec [▼] [Editar...] [Nuevo...] [Wizard...]

SQL

```
SELECT * FROM usuarios  
WHERE activo = true
```

Obtener consulta SQL...

Line 2 Column 19

Activar conversión perezosa. ☐

¿Reemplazar variables en script? ☐

Insertar datos del paso

¿Ejecutar para cada fila? ☐

Limitar tamaño: 0

[Help] [OK] [Previsualizar] [Cancelar]

Database Connection

General

Connection Name:

Connection Type:

- LucidDB
- MS Access
- MS SQL Server
- MS SQL Server (Native)
- MaxDB (SAP DB)
- MonetDB
- MySQL
- Native Mondrian
- Neoview
- Netezza
- OpenERP Server
- Oracle
- Access:
- Native (JDBC)
- ODBC
- OCI
- JNDI

Settings:

Host Name:

Database Name:

Tablespace for Data:

Tablespace for Indices:

Port Number: 1521

User Name:

Password:

[Probar] [Lista de f.] [Explorar]

[OK] [Cancelar]

PDI Pasos de ENTRADA

Lectura de Archivo



Entrada Fichero de Texto

Archivo de texto - Entrada

Nombre de paso: Entrada Fichero de Texto

Fichero | Contenido | Manejo de Errores | Filtros | Campos | Additional output fields

Tipo de fichero: CSV

Separador de campos: ,

Separador de texto: "

de linea en campos con separador de texto? ☐

Escape:

Cabecera ☒ Número de líneas de cabecera: 1

Pie ☐ Número de líneas de pie: 1

¿Líneas cortadas? ☐ Número de veces que se corta: 1

Paginado (impresión)? ☐ Número de líneas por página: 80

Líneas en cabecera documento:

Comprimido (Zip): None

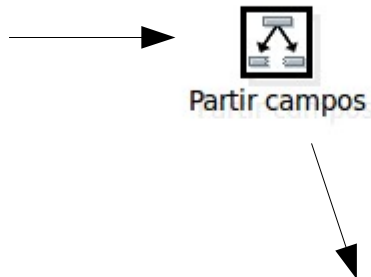
Eliminar filas vacías ☒

Help OK Previsualizar filas Cancelar

PDI Pasos de TRANSF

PARTIR CAMPOS

razon_soc
ROJAS, Ariel
MORA, Rodrigo



razon_soc	Nombre	Apellido
ROJAS, Ariel	Ariel	ROJAS
MORA, Rodrigo	Rodrigo	MORA

PDI Pasos de TRANSF

OPERACIONES SOBRE STRINGS

Localidad	cod
La plata	29
Buenos aires	0021



Localidad	cod
La Plata	00029
Buenos Aires	00021

PDI Pasos de TRANSF SCRIPTING

fecha
1-1-2014
2-1-2014



Valor Java Script Modificado

fecha	trimestre	anio	mes	semana
1-1-2014	1	2014	1	1
2-1-2014	1	2014	1	1

Valores de Script

Nombre de paso: Valor Java Script Modificado

Funciones Javascript:script :

- truncD
- week(v
- year(v
- Logic Fun
- Special Fu
- File Funct

Script 1

```
var trimestre = quarter(fecha);  
var anio = year(fecha);  
var mes = month(fecha);  
var semana = week(fecha);
```

Posición: 1

¿Modo de c: ☐ Optimization level: 9

Campos

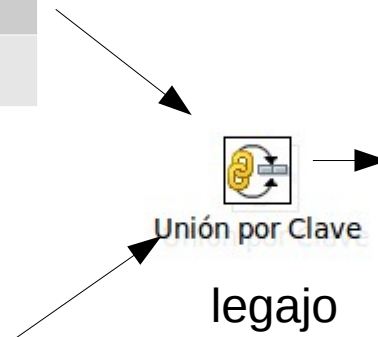
#	Nombre de campo	Renombar a	Tipo
1	trimestre		
2	anio		
3	mes		
4	semana		

OK Cancelar Obtener Variables Probar scr

PDI Pasos de TRANSF

UNION POR CLAVE

legajo	nombre
1900	Teofilo Gutierrez
1901	Matías Kranevitter



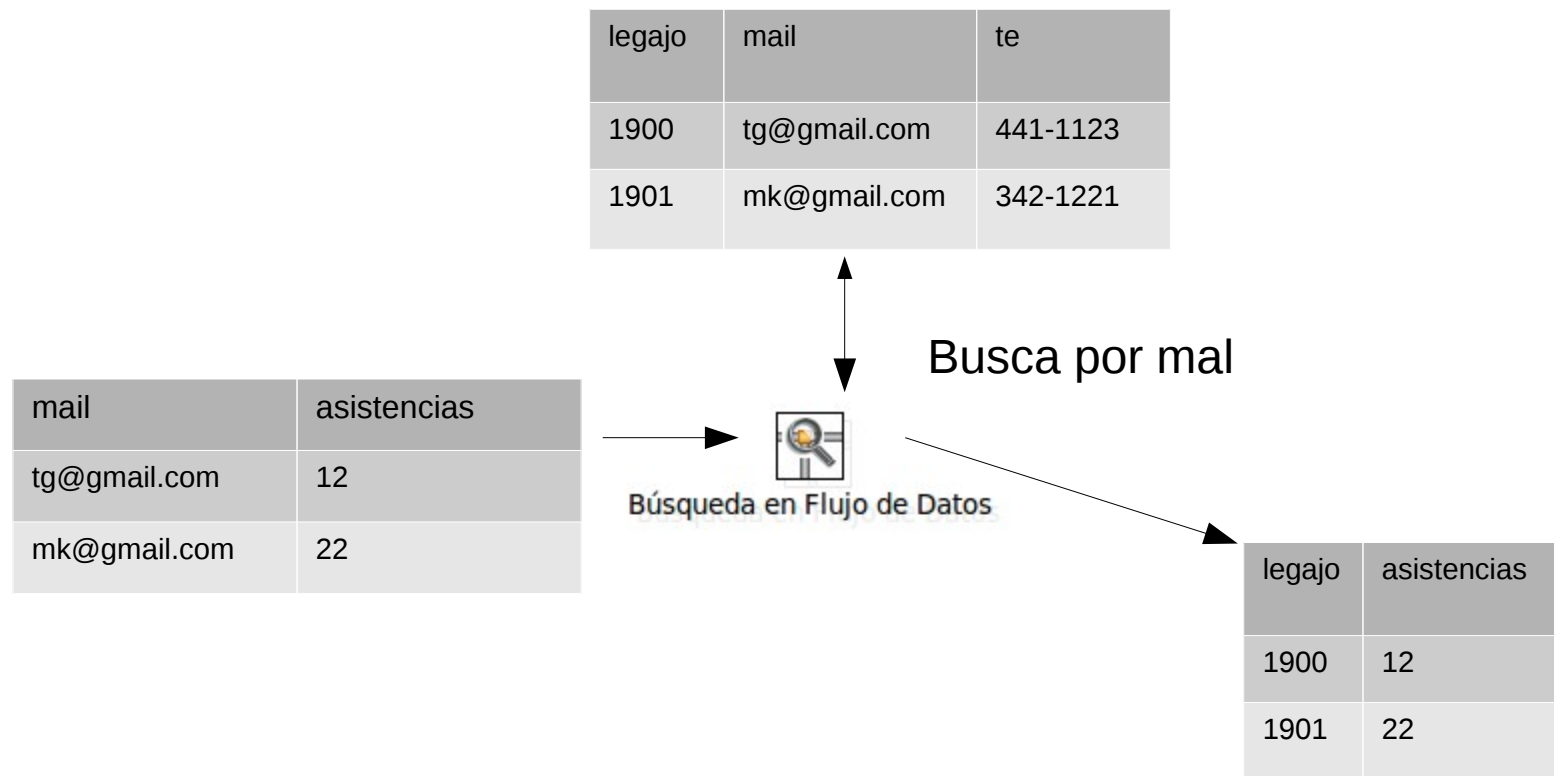
legaj o	nombre	mail	te
1900	Teofilo Gutierrez	tg@gmail.com	441-1123
1901	Matias Kranevitter	mk@gmail.com	342-1221

legajo	mail	te
1900	tg@gmail.com	441-1123
1901	mk@gmail.com	342-1221

Algunos pasos, como este,
requieren la entrada ordenada

PDI Pasos de TRANSF

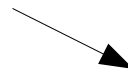
BUSQUEDA EN FLUJO DE DATOS



PDI Pasos de TRANSF

SELECCIONA / RENOMBRA

c1	c2	c3	c4	c5	c6



Selecciona/Renombra valores

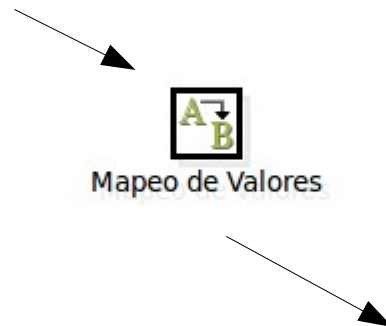


c5	c3	c1	c2

PDI Pasos de TRANSF

MAPEO DE VALORES

Genero	estudio	trabaja
Masc	pri	Si
Fem	Primario	False
M	sec	True
M	2rio	1
F	univ	0



Genero	estudio	trabaja
Masculino	Primario	Trabaja
Femenino	Primario	No Trabaja
Masculino	Secundario	Trabaja
Masculino	Secundario	Trabaja
Femenino	Universitario	No Trabaja

PDI Pasos de TRANSF

ORDENA FILAS

dni	nombre
23456789	Marquezi, Jose
21234321	Rodriguez, Jorge



Este paso requiere obtener
TODAS las filas antes de continuar

dni	nombre
21234321	Rodriguez, Jorge
23456789	Marquez, Jose

PDI Pasos de TRANSF

FLUJO : FILTRAR FILAS

num
1
2
3
4
5
6
7

CSV file input

Filtrar filas (es par?)

pares

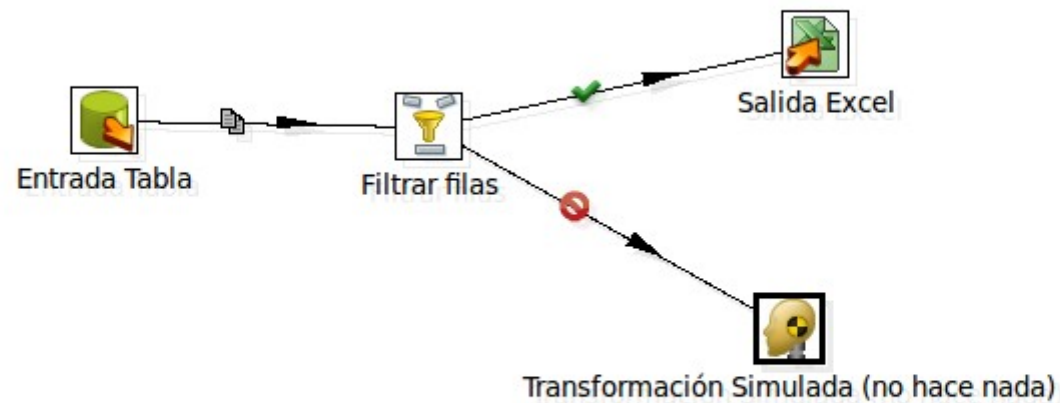
impares

num
2
4
6

num
1
3
5
7

PDI Pasos de TRANSF

FLUJO : PASO NO HACE NADA



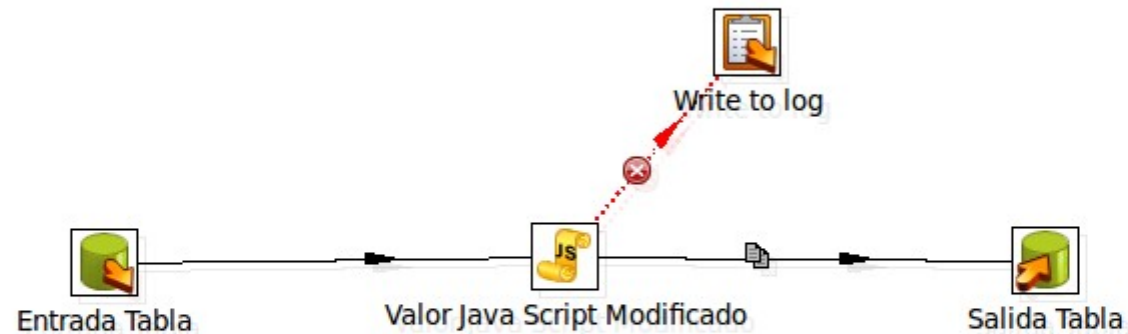
PDI Pasos de TRANSF

FLUJO : COPIA vs DISTRIBUYE



PDI Pasos de TRANSF

FLUJO : MANEJO DE ERRORES



PDI Pasos de TRANSF

DWH – Actualizar Dimension

The screenshot shows the PDI (Pentaho Data Integration) interface. On the left, a data flow diagram shows a source table 'tipo_inmueble' being transformed into a dimension 'dim_tipo_inmueble'. On the right, the 'Búsqueda/Actualización de Dimensión' (Dimension Search/Update) dialog is open for the step 'dim_tipo_inmueble'. The dialog has the following settings:

- Nombre del paso: dim_tipo_inmueble
- ¿Actualizar la dimensión?: ☒
- Conexión: dwh
- Esquema destino: (empty)
- Tabla destino: dim_tipo_inmueble
- Tamaño de transacción: 100
- ¿Habilitar la cache?: ☒
- ¿Precargar la cache?: ☐
- Tamaño del cache, en filas (0 = todo): 5000

Below these settings are two tabs: 'Claves' (Keys) and 'Campos' (Fields). The 'Claves' tab shows a single key: 'id'. The 'Campos' tab shows a table for mapping fields from the source flow to the dimension:

#	Campo de Dimensión	Campo en el flujo
1	id	id

Below this is another section titled 'Campos de Búsqueda/Actualización' (Search/Update Fields) with a table:

#	Campo de Dimensión	Campo del flujo con el c	Tipo de actualización de c
1	descripcion	descripcion	Punch through (actualizar
2	grupo	grupo	Actualizar
3	valor_m2	valor	Insertar

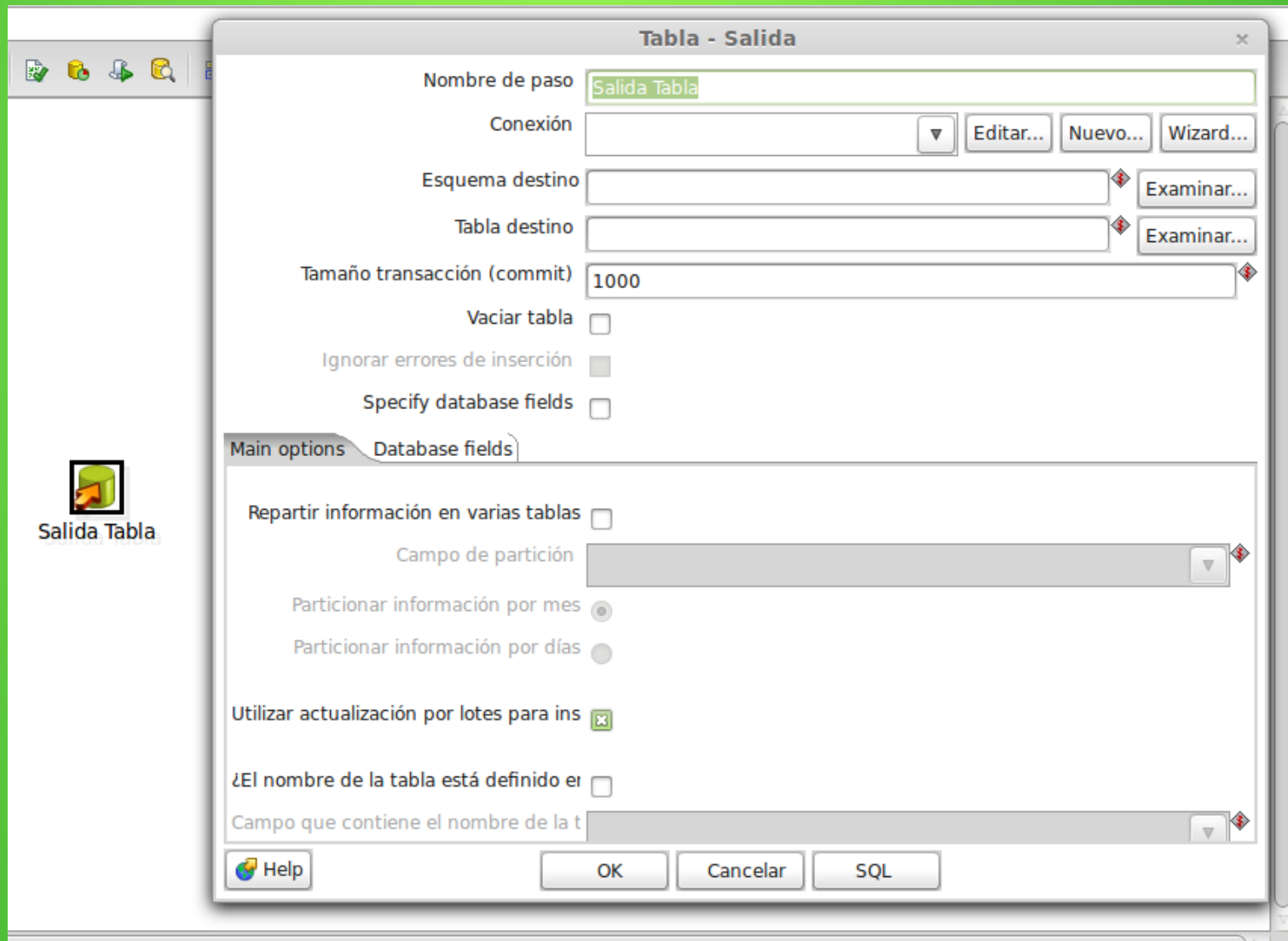
Three blue callout boxes point to specific elements in the 'Campos' tab:

- 'Campo de búsqueda' points to the 'id' key in the 'Claves' tab.
- 'Actualiza Todas las versiones' points to the 'Punch through (actualizar' update type for 'descripcion'.
- 'Actualiza la Ultima versión' points to the 'Actualizar' update type for 'grupo'.
- 'Inserta una nueva versión del registro' points to the 'Insertar' update type for 'valor_m2'.

pk	id	descripcion	grupo	valor_m2	date_from	date_to	version
1	1	Residencial	Edificado	821,00	1-1-1900	21-9-2014	1
2	1	Residencial	Edificado	910,00	21-9-2014	31-12-2100	2

PDI Pasos de SALIDA

SALIDA A TABLA



PDI Pasos de SALIDA

INSERTAR / ACTUALIZAR

Insertar / Actualizar

Nombre del paso: Insertar / Actualizar

Conexión: [dropdown] [Editar...] [Nuevo...] [Wizard...]

Esquema destino: [dropdown] [Examinar...]

Tabla destino: tabla de búsqueda [Examinar...]

Tamaño de transacción (compr): 100

No realizar actualizaciones: ☐

La(s) clave(s) para realizar la búsqueda de valor(es):

#	Campo de tabla	Comparador	Campo de flujo1
1	ID	=	CODIGO

[Obtener campos]

Campos de actualización:

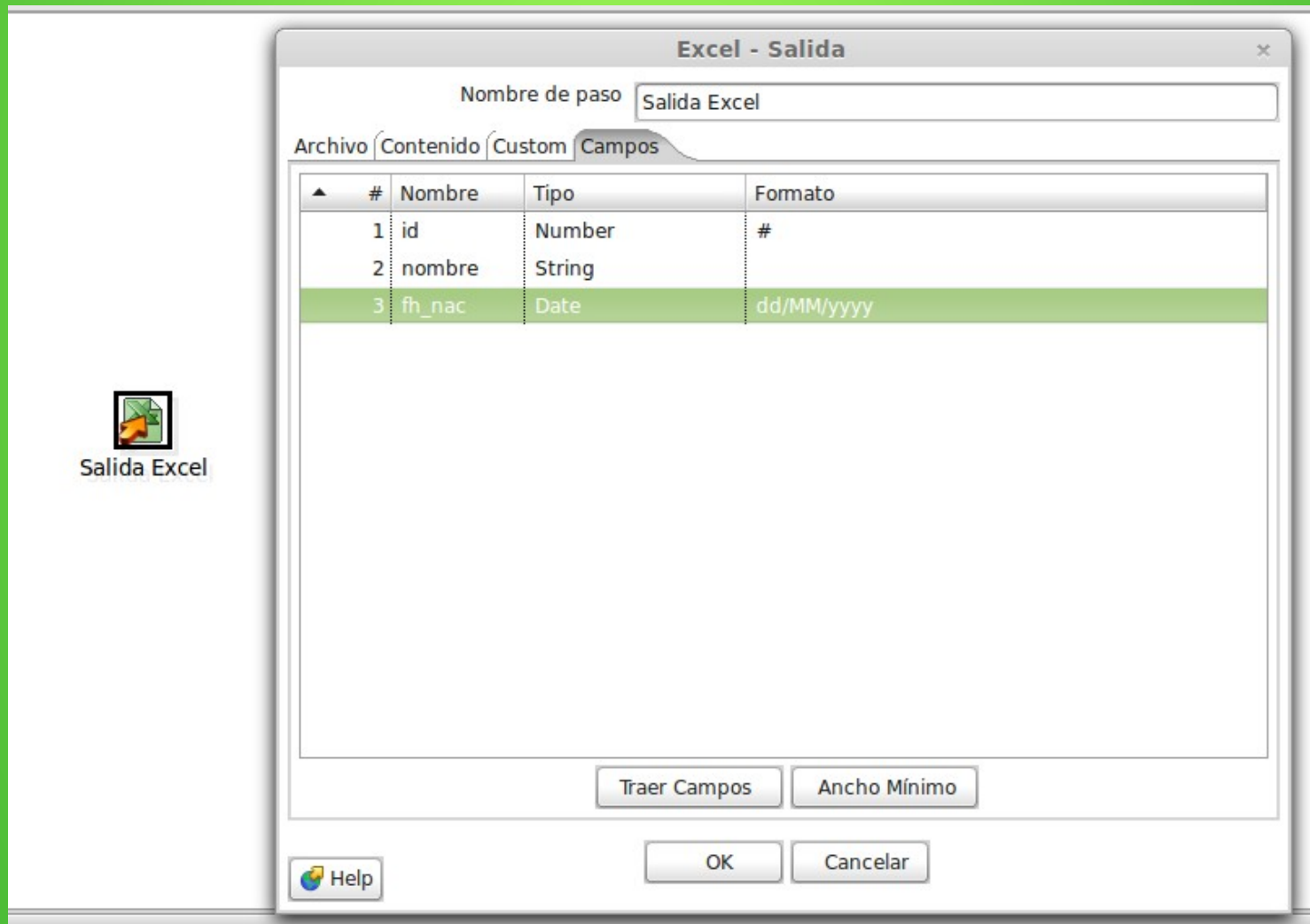
#	Campo de tabla	Campo de Flujo
1	ID	CODIGO
2	NOMBRE	NOM
3	APELLIDO	AP

[Obtener campos de actualización] [Editar mapeo]

[Help] [OK] [Cancelar] [SQL]

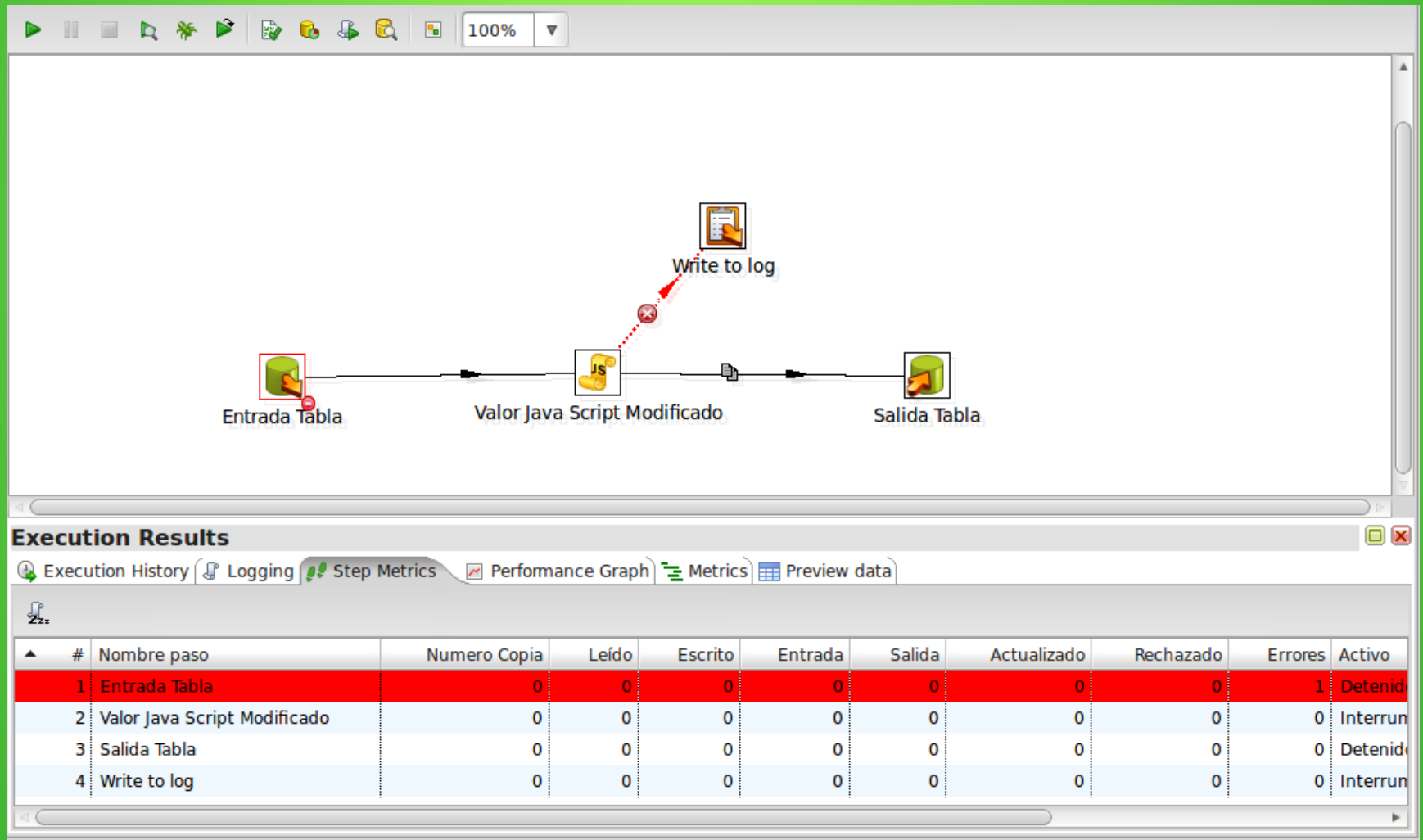
PDI Pasos de SALIDA

SALIDA A EXCEL



PDI EJECUCION

entrada - salida - leído - escrito



PDI EJECUCION

LOG DE ERRORES

The screenshot displays the Pentaho Data Integration (PDI) user interface. At the top, a toolbar contains various icons for job execution and management, along with a zoom level set to 100%. The main workspace shows a job graph with three steps: 'Entrada Tabla' (Table Input), 'Valor Java Script Modificado' (Modified JavaScript Value), and 'Salida Tabla' (Table Output). A 'Write to log' step is connected to the 'Valor Java Script Modificado' step via a red dashed line, indicating an error. Below the workspace, the 'Execution Results' panel is visible, showing tabs for 'Execution History', 'Logging', 'Step Metrics', 'Performance Graph', 'Metrics', and 'Preview data'. The 'Logging' tab is selected, displaying a log of execution events. The log shows several error messages from the Spoon engine, indicating that errors were detected during the execution of the job.

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

2014/09/22 14:28:19 - Spoon -
2014/09/22 14:28:19 - Spoon -
2014/09/22 14:28:19 - Spoon - at org.pentaho.di.trans.Trans.prepareExecution(Trans.java:1068)
2014/09/22 14:28:19 - Spoon - at org.pentaho.di.ui.spoon.trans.TransGraph\$26.run(TransGraph.java:3885)
2014/09/22 14:28:19 - Spoon - at java.lang.Thread.run(Thread.java:744)
2014/09/22 14:28:19 - AAA - ERROR (version 5.0.1-stable, build 1 from 2013-11-15_16-08-58 by buildguy) : ¡Errores detectados!
2014/09/22 14:28:19 - AAA - ERROR (version 5.0.1-stable, build 1 from 2013-11-15_16-08-58 by buildguy) : ¡Errores detectados!

PDI EJECUCION

PARAMETROS

Ejecutar una transformación

Ejecución local, remota o clustered

☒ Ejecución local ☐ Ejecución remota ☐ Ejecución clustered

Servidor remoto

☐ Pass export to remote server

☒ Enviar transformación
☒ Preparar ejecución
☒ Iniciar ejecución
☐ Mostrar transformaciones

Detalles

☐ Habilitar modo seguro
☒ Gather performance metrics
☒ Clear the log before execution

Nivel de registro

Fecha de Ejecución (yyyy/MM/dd HH:mm:ss)

Parameters

#	Parameter	Value	Default value
1	periodo		

Parámetros

#	Parámetro	Valor
1		

Variables

#	Variable	Valor
1	Internal.Job.Filename.Directory	Parent Job File Dire
2	Internal.Job.Filename.Name	Parent Job Filename
3	Internal.Job.Name	Parent Job Name
4	Internal.Job.Repository.Directory	Parent Job Reposit

PDI EJECUCION

PARAMETROS



Tabla - Entrada

Nombre paso: Entrada Tabla

Conexión: [▼] [Editar...] [Nuevo...] [Wizard...]

SQL: [Obtener consulta SQL...]

`SELECT * FROM data WHERE periodo=?`

Line 1 Column 34

Activar conversión perezosa. ☐

¿Reemplazar variables en script? ☐

Insertar datos del paso: [Obtener Variables ▼]

¿Ejecutar para cada fila? ☐

Limitar tamaño: [0 ▼]

[Help] [OK] [Previsualizar] [Cancelar]

PDI EJECUCION

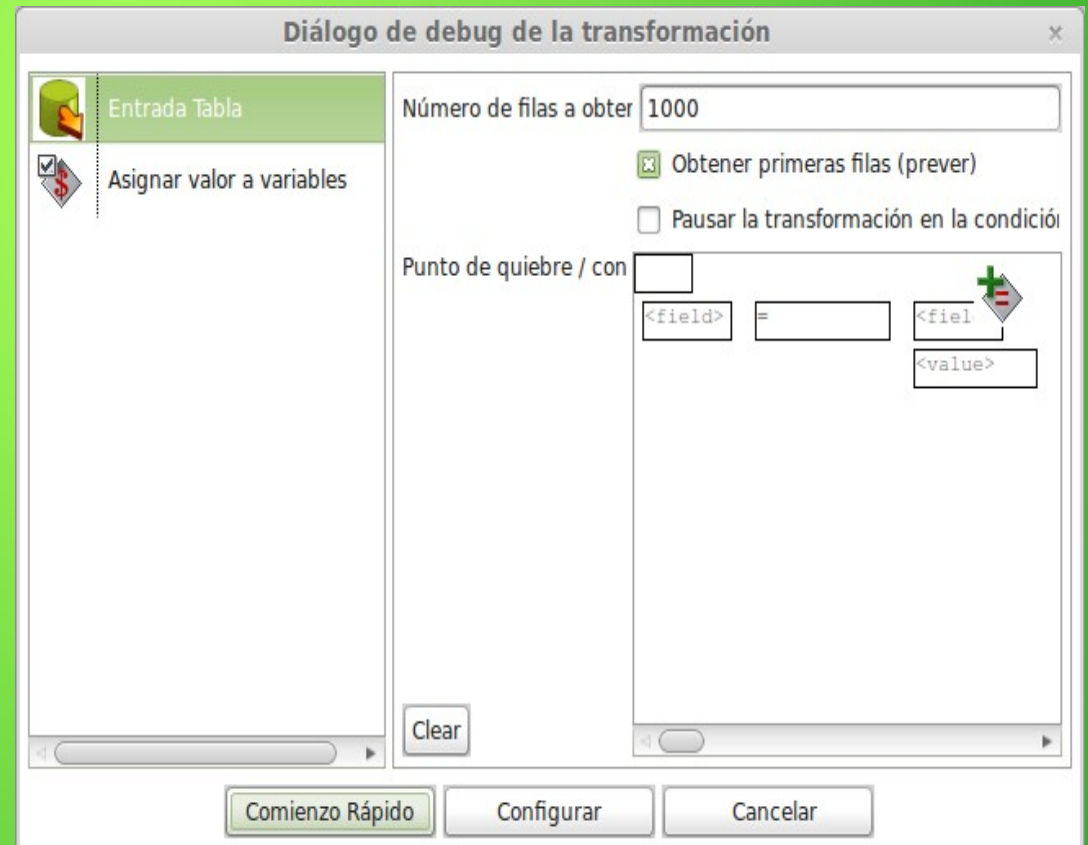
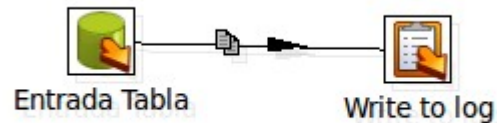
DEBUG

PREVISUALIZAR

VER PASOS DE ENTRADA

VER PASOS DE SALIDA

ESCRIBIR AL LOG

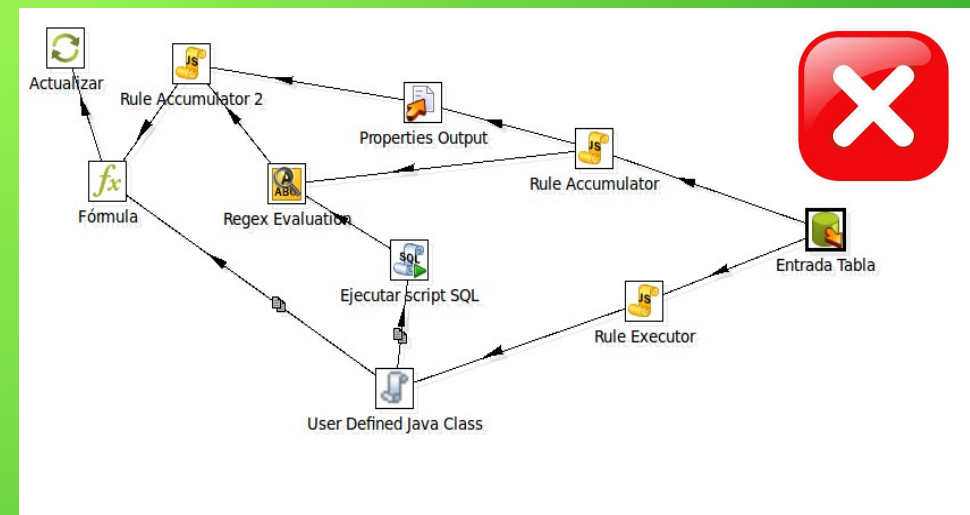
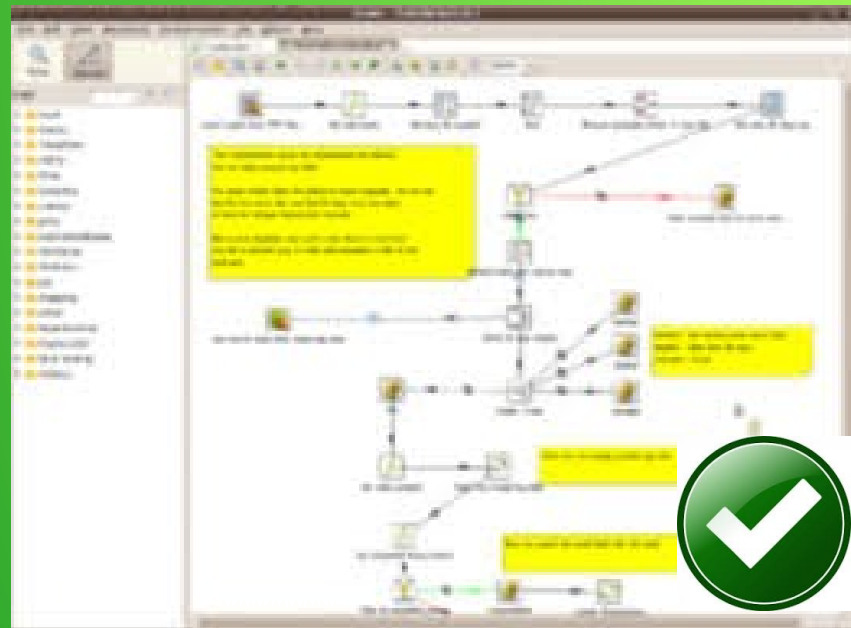
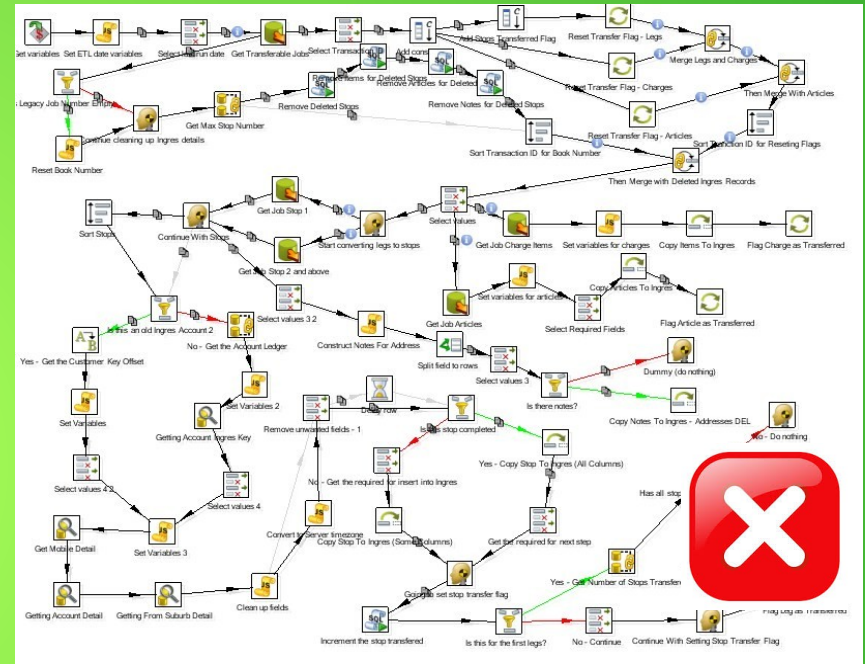
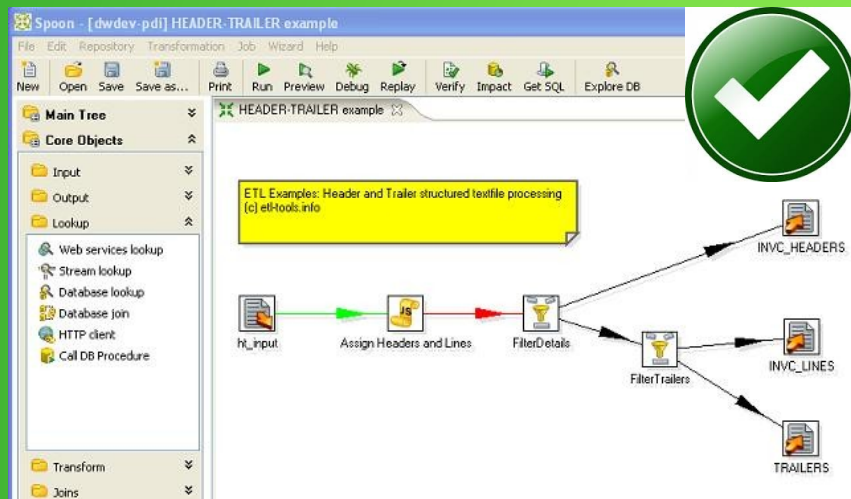


PDI

CONSEJOS UTILES

- Flujo de Izq → Der | Arriba → Abajo
- Utilizar nombres descriptivos para los pasos
- Documentar con Notas
- Mantener los archivos simples
- Separar en varias transformaciones – jobs para simplificar
- Utilizar nombres vinculos relativos en lugar de absolutos
- Organizar los archivos para que se correspondan con los nombres de los pasos
- Identificar facilmente el trabajo inicial (main.kbj)

PDI - Ejemplos

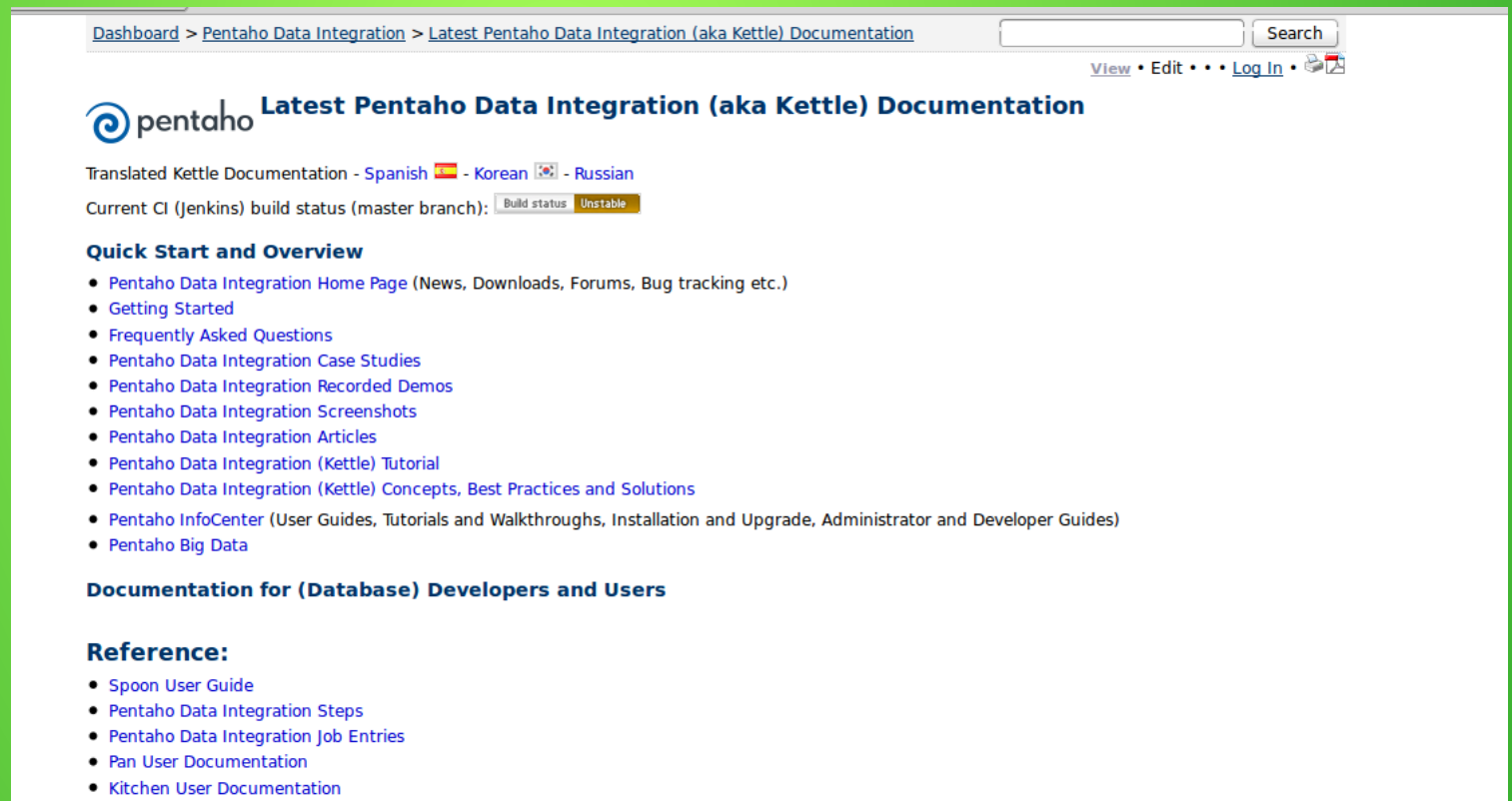


PDI

LINKS

DOCUMENTACION

<http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation>



The screenshot shows the Pentaho Data Integration (aka Kettle) Documentation page. The page has a navigation bar at the top with links: Dashboard > Pentaho Data Integration > Latest Pentaho Data Integration (aka Kettle) Documentation. There is a search bar and a 'Search' button. Below the navigation bar, the page title is 'Latest Pentaho Data Integration (aka Kettle) Documentation'. The Pentaho logo is on the left. Below the title, there are links for 'Translated Kettle Documentation - Spanish', 'Korean', and 'Russian'. A status bar shows 'Current CI (Jenkins) build status (master branch): Build status Unstable'. The main content area is divided into sections: 'Quick Start and Overview' with a list of links (Pentaho Data Integration Home Page, Getting Started, Frequently Asked Questions, Pentaho Data Integration Case Studies, Pentaho Data Integration Recorded Demos, Pentaho Data Integration Screenshots, Pentaho Data Integration Articles, Pentaho Data Integration (Kettle) Tutorial, Pentaho Data Integration (Kettle) Concepts, Best Practices and Solutions, Pentaho InfoCenter, Pentaho Big Data), 'Documentation for (Database) Developers and Users', and 'Reference:' with a list of links (Spoon User Guide, Pentaho Data Integration Steps, Pentaho Data Integration Job Entries, Pan User Documentation, Kitchen User Documentation).

Dashboard > Pentaho Data Integration > Latest Pentaho Data Integration (aka Kettle) Documentation

Search

View • Edit • • • Log In •

pentaho Latest Pentaho Data Integration (aka Kettle) Documentation

Translated Kettle Documentation - Spanish - Korean - Russian

Current CI (Jenkins) build status (master branch): Build status Unstable

Quick Start and Overview

- [Pentaho Data Integration Home Page](#) (News, Downloads, Forums, Bug tracking etc.)
- [Getting Started](#)
- [Frequently Asked Questions](#)
- [Pentaho Data Integration Case Studies](#)
- [Pentaho Data Integration Recorded Demos](#)
- [Pentaho Data Integration Screenshots](#)
- [Pentaho Data Integration Articles](#)
- [Pentaho Data Integration \(Kettle\) Tutorial](#)
- [Pentaho Data Integration \(Kettle\) Concepts, Best Practices and Solutions](#)
- [Pentaho InfoCenter](#) (User Guides, Tutorials and Walkthroughs, Installation and Upgrade, Administrator and Developer Guides)
- [Pentaho Big Data](#)

Documentation for (Database) Developers and Users

Reference:

- [Spoon User Guide](#)
- [Pentaho Data Integration Steps](#)
- [Pentaho Data Integration Job Entries](#)
- [Pan User Documentation](#)
- [Kitchen User Documentation](#)

DESCARGA

<http://community.pentaho.com/projects/data-integration/>

PDI

?

PDI

DEMOSTRACION

DIMENSIONES

- Barrio
- Tipo Inmueble
- Estado Inmueble
- Estado Parcela

MEDIDAS

- Cant Parcelas
- Cant Inmuebles
- Superficie parcela (m2)

	dim_estado_constructivo
	descripcion

	dim_tipo_inmueble
	descripcion

	fact_parcelas
	superficie_m2
	cant_parcelas
	cant_inmuebles

	dim_tipo_parcela
	descripcion

	dim_barrio
	nombre
	descripcion(A)

PDI

DEMOSTRACION

