

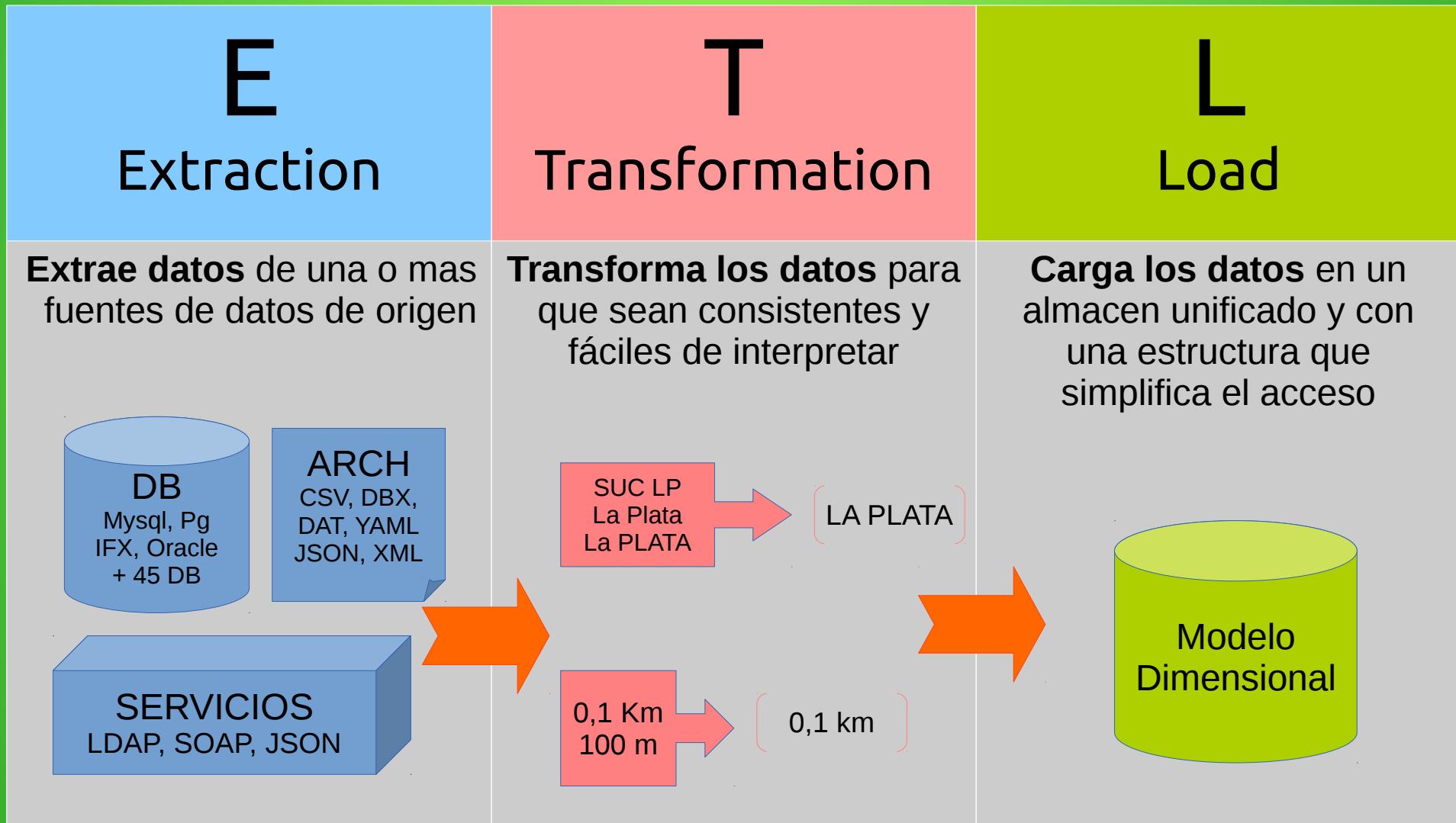
# Pentaho Data Integration

PDI

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

Se utiliza para poblar DWH  
Migrar Datos  
Exportar vistas

# ETL



# PDI

JAVA - Multiplataforma

**EJECUTAR DETENER**

**AREA DE TRABAJO**

**INFO EJECUCION**

The screenshot shows the Spoon interface for PDI. On the left, the 'Steps' palette lists various data sources and transformations. The main workspace displays a flow starting with 'Read Sales Data', followed by 'Filter Missing Zips', 'Value Mapper', 'Select values', 'Number range', and finally 'Write to Database'. A 'Prepare Field Layout' step is also present, connected to 'Lookup Missing Zips'. A tooltip provides instructions for testing the transformation. The bottom section shows the 'Execution Results' table with detailed statistics for each step.

Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	Inp
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		

PASOS

# PDI

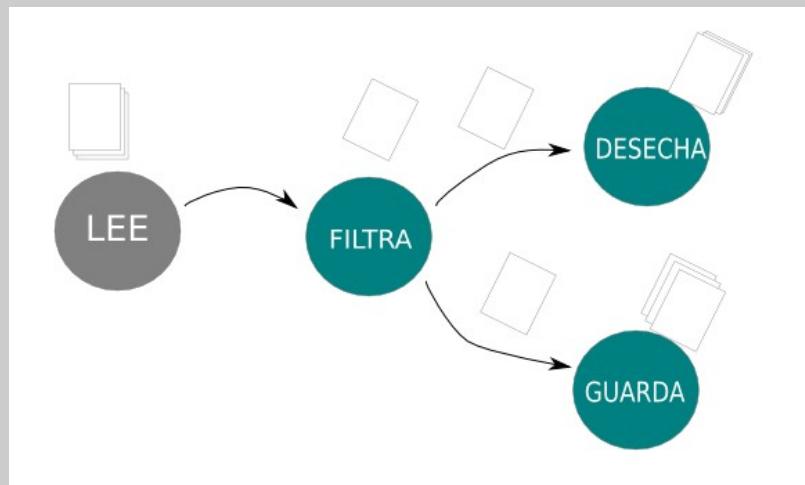
## Transformaciones

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

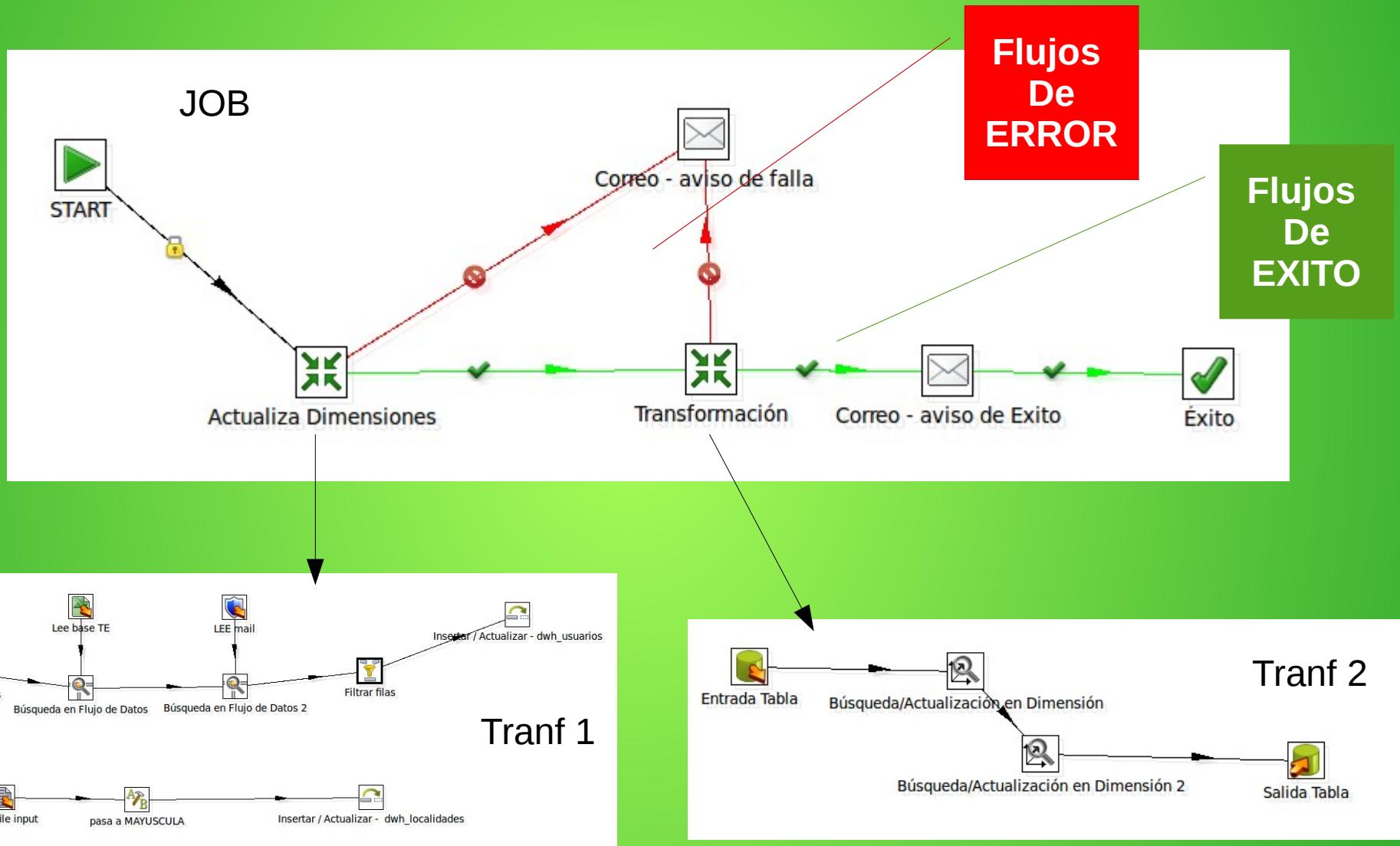
Kettle Job \*.kjb

Se ejecuta de un paso a la vez.

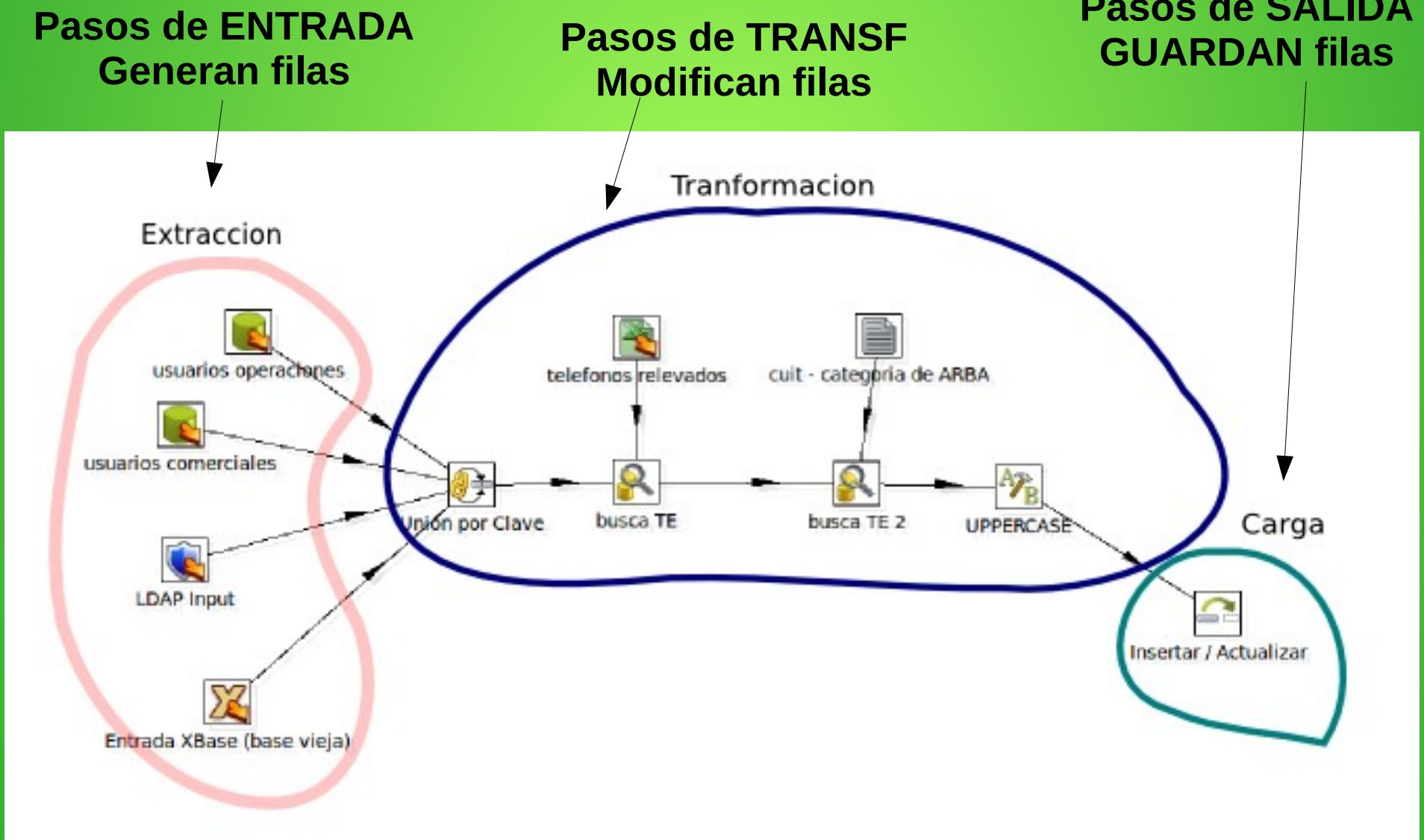
Se utiliza para organizar varias transformaciones



# PDI



# PDI



# PDI Pasos de ENTRADA

## Consulta a Base de Datos



Entrada Tabla

**Tabla - Entrada**

Nombre paso: Entrada Tabla  
Conexión: guarani.ol\_agro\_sec

SQL:

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

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 Cancel

**Database Connection**

General Advanced Options Pooling Clustering

Connection Name:

Connection Type: Oracle

Settings: Host Name:   
Database Name:   
Tablespace for Data:   
Tablespace for Indices:   
Port Number: 1521  
User Name:   
Password:

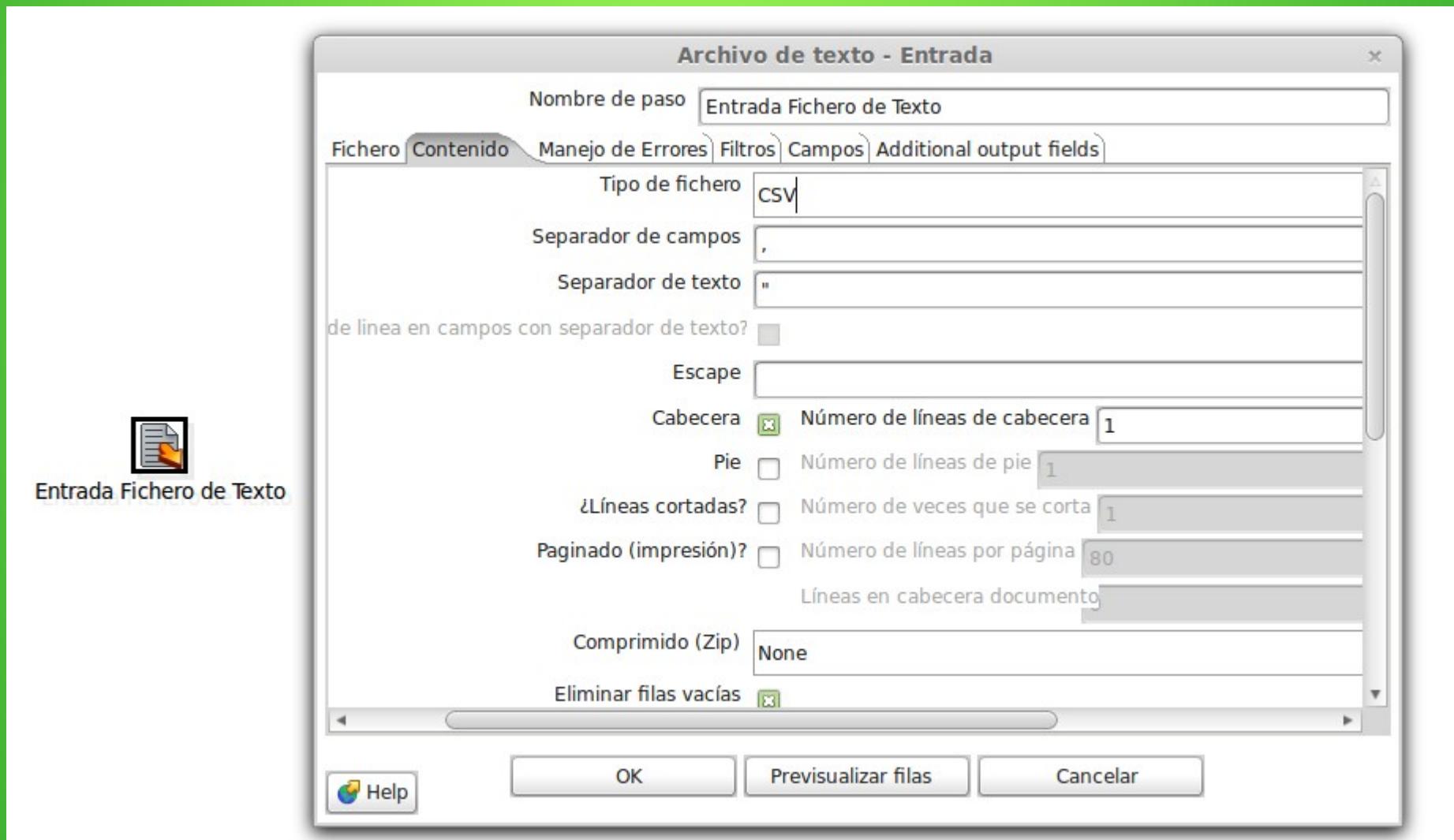
Access: Native (JDBC)

ODBC OCI JNDI

Probar Lista de fuentes Explorar OK Cancel

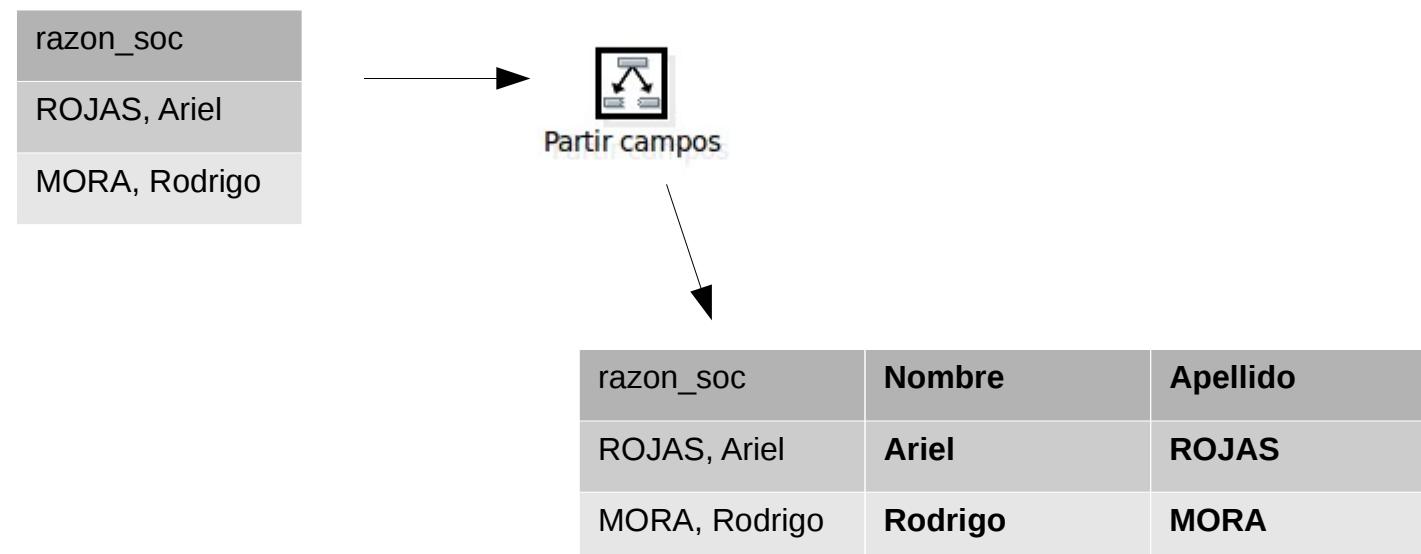
# PDI Pasos de ENTRADA

## Lectura de Archivo



# PDI Pasos de TRANSF

## PARTIR CAMPOS



# 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

The diagram illustrates a data transformation process using PDI (Apache Nifi). On the left, a table shows two rows of input data:

fecha				
1-1-2014				
2-1-2014				

An arrow points from the 'fecha' column of the first row to a 'JS' icon, which is labeled 'Valor Java Script Modificado'. This icon points down to a second table showing the resulting transformed data:

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

To the right, a 'Valores de Script' dialog box is shown, containing the following details:

- Nombre de paso:** Valor Java Script Modificado
- Funciones Javascript:script :**
  - truncDate
  - week
  - year

Logic Functions

Special Functions

File Functions
- 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		

**Buttons:** OK, Cancelar, Obtener Variables, Probar scr

# PDI Pasos de TRANSF

## UNION POR CLAVE

legajo	nombre
1900	Teofilo Gutierrez
1901	Matías Kranevitter



Unión por Clave

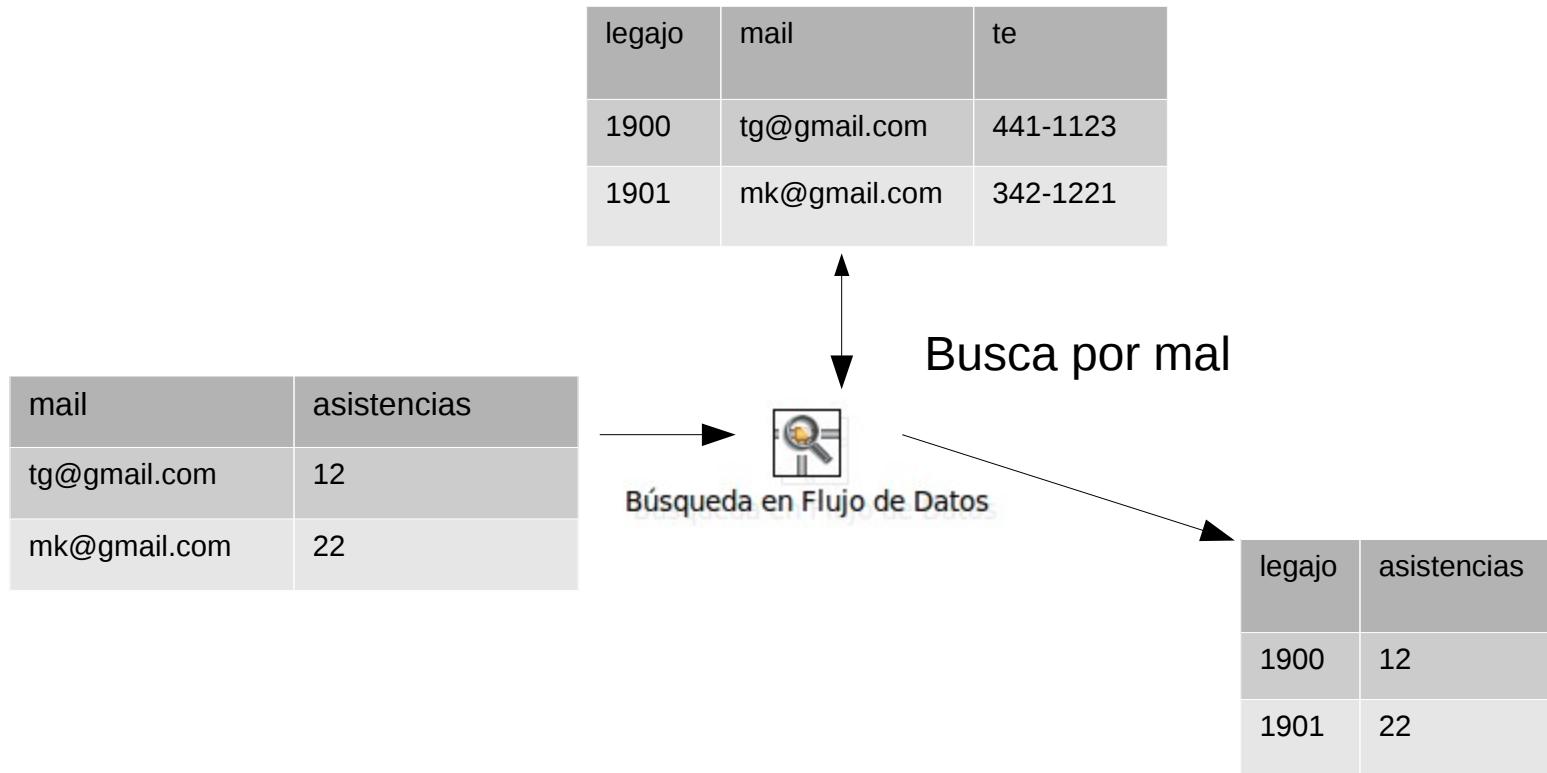
legajo	nombre	mail	te
1900	Teofilo Gutierrez	tg@gmail.com	441-1123
1901	Matias Kranevitter	mk@gmail.com	342-1221

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

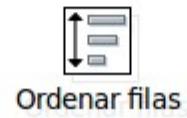


Genero	estudio	trabaja
Masculino	Primario	Trabaja
Femenino	Primario	No Trabaja
Masculino	Secundario	Trabaja
Masculino	Secundario	Trabaja
Femenino	Universitario	No Trabaja

# PDI Pasos de TRANSF

## ORDENA FILAS

dni	nombre
23456789	Marquezi, Jose
21234321	Rodriguez, Jorge



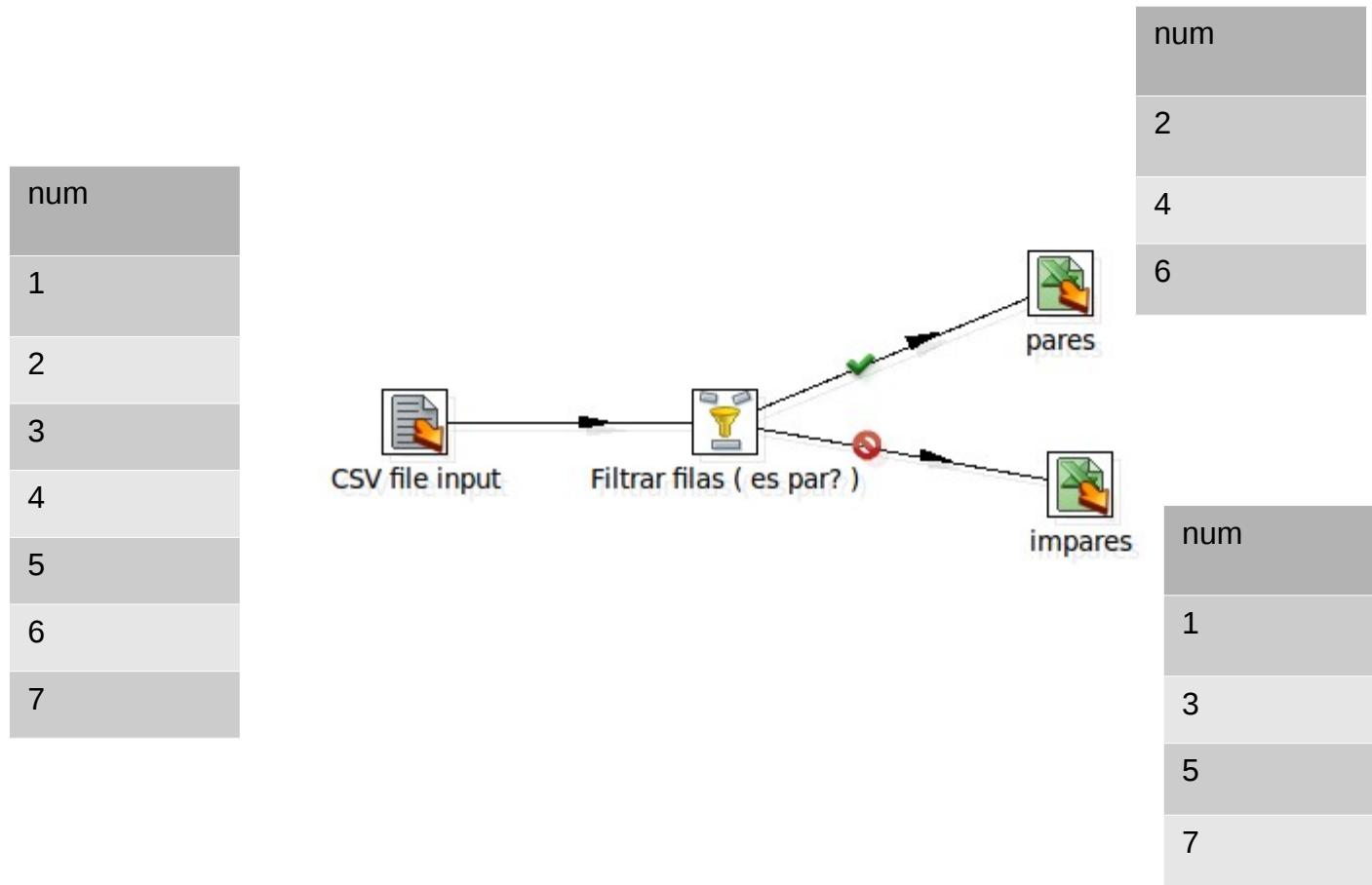
Ordenar filas

Este paso requiere obtener  
TODAS las filas antes de continuar

dni	nombre
21234321	Rodriguez, Jorge
23456789	Marquez, Jose

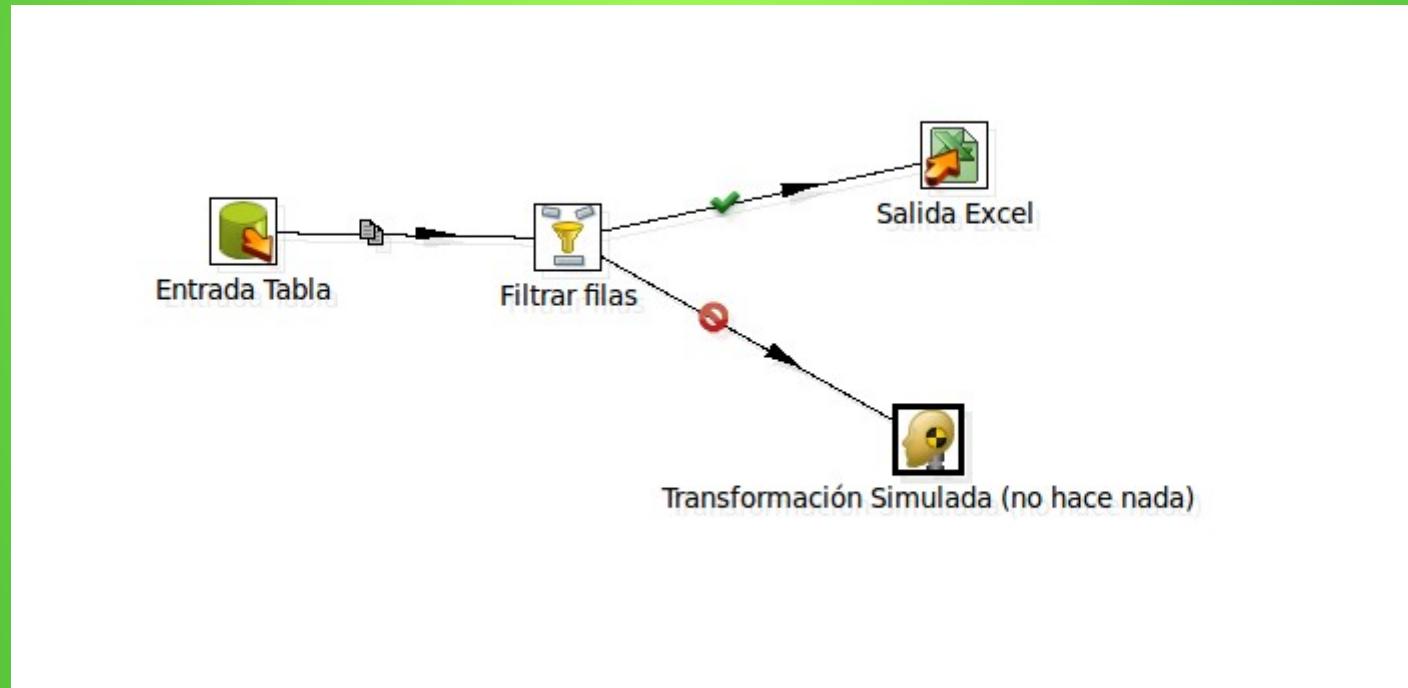
# PDI Pasos de TRANSF

## FLUJO : FILTRAR FILAS



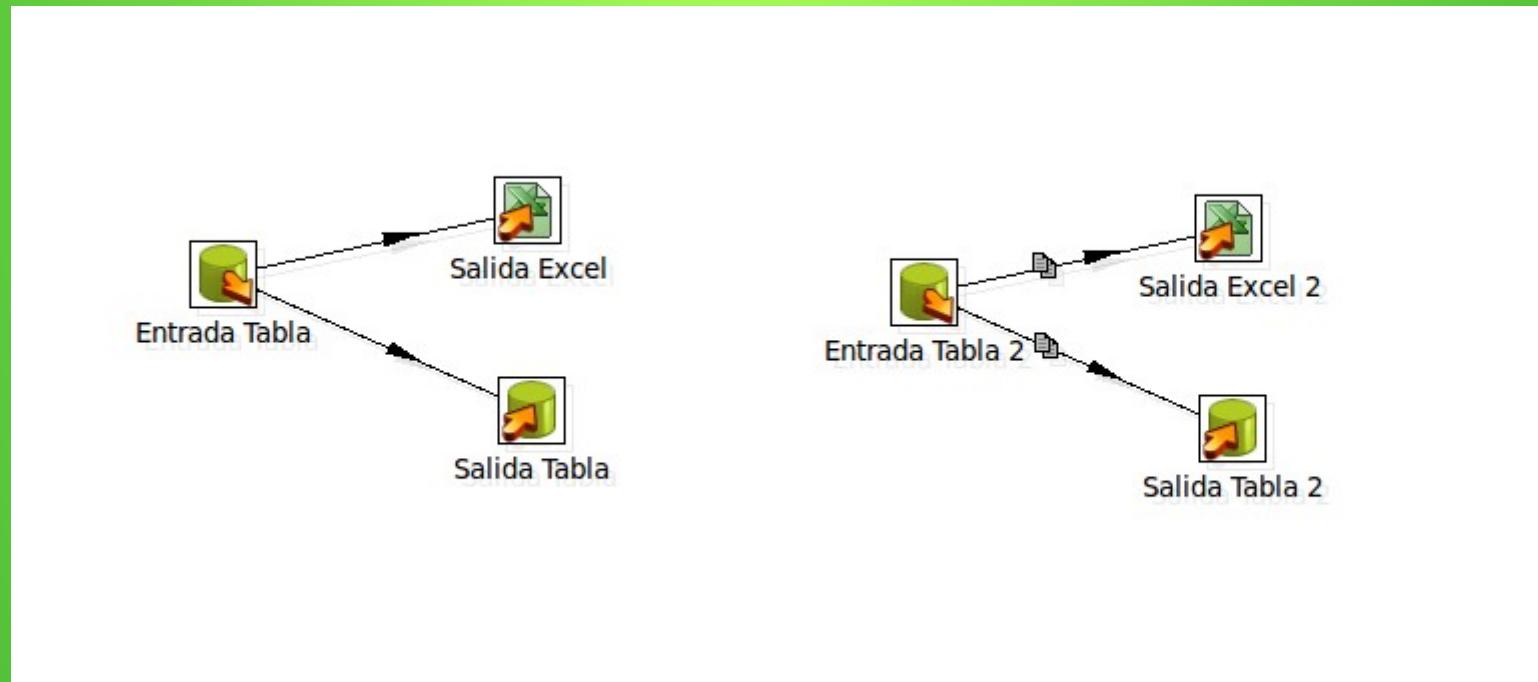
# PDI Pasos de TRANSF

FLUJO : PASO NO HACE NADA



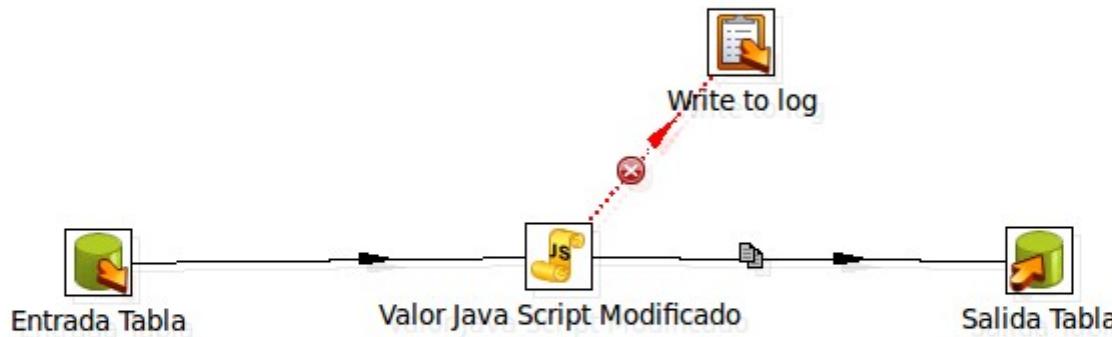
# PDI Pasos de TRANSF

FLUJO : COPIA vs DISTRIBUYE



# PDI Pasos de TRANSF

## FLUJO : MANEJO DE ERRORES



# PDI Pasos de TRANSF

## DWH – Actualizar Dimension

The screenshot shows the PDI Data Integration interface. On the left, a data flow diagram displays four parallel steps: 'estado\_constructivo' feeds into 'dim\_estado\_constru', 'tipo\_inmueble' feeds into 'dim\_tipo\_inmueble', 'estado\_parcela' feeds into 'dim\_estado\_parcela', and 'barrio' feeds into 'dim\_barrio'. Each step consists of a green cylinder icon followed by a magnifying glass icon. On the right, a configuration window titled 'Búsqueda/Actualización de Dimensión' is open. It contains the following settings:

- Nombre del paso: dim\_tipo\_inmueble
- ¿Actualizar la dimensión?:
- Conexión: dwh
- Esquema destino:
- 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' and 'Campos'. The 'Claves' tab shows a single key mapping: 'id' from the dimension field to 'id' in the flow field. The 'Campos' tab lists fields for search/update:

#	Campo de Dimensión	Campo en el flujo	Tipo de actualización de campo
1	descripcion	descripcion	Punch through (actualiza)
2	grupo	grupo	Actualizar
3	valor_m2	valor	Insertar

On the far right, four blue callout boxes provide additional context:

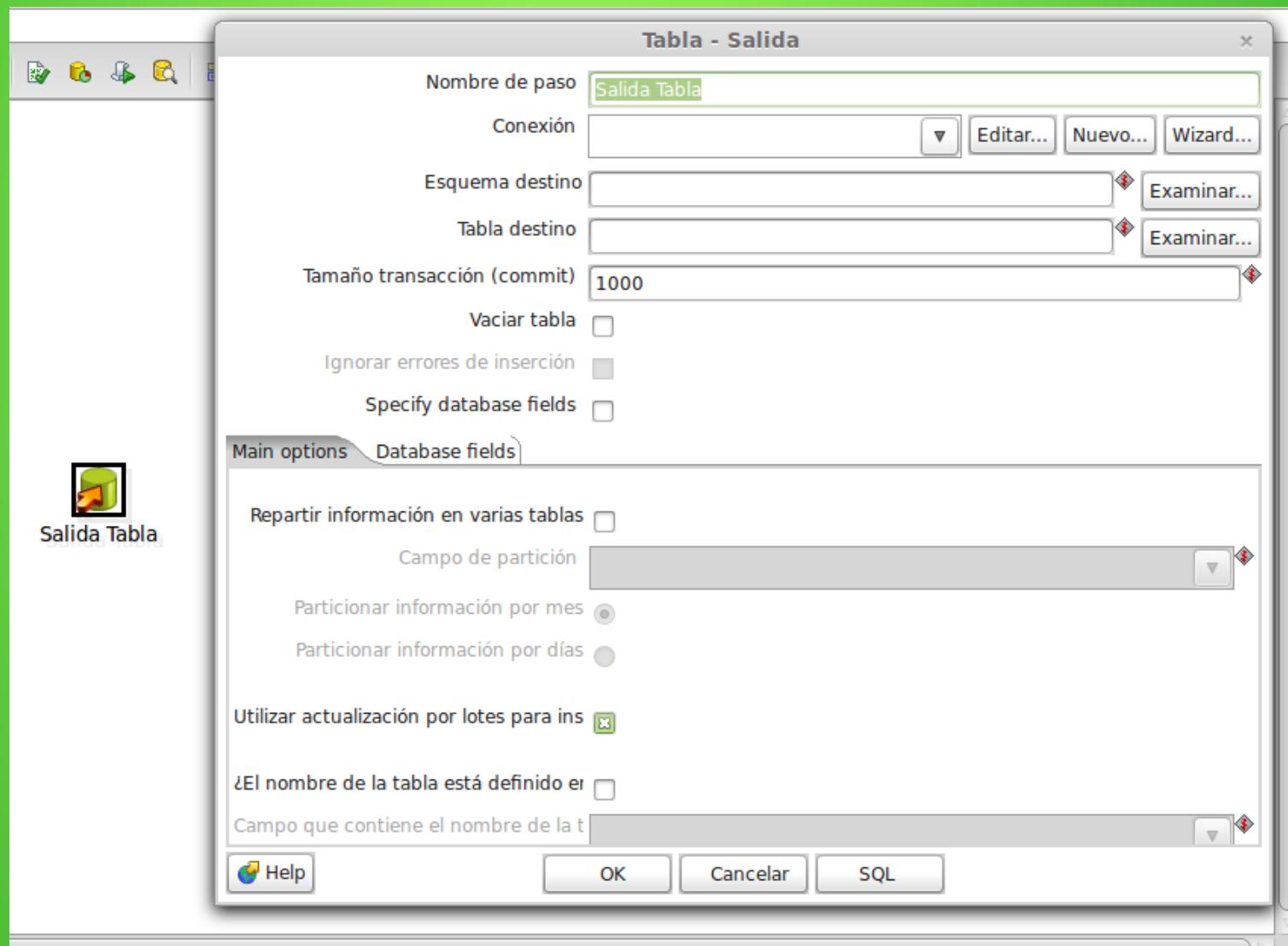
- Campo de búsqueda
- Actualiza Todas las versiones
- Actualiza la Última versión
- Inserta una nueva versión del registro

At the bottom, a table shows the current state of the 'dim\_tipo\_inmueble' dimension table:

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

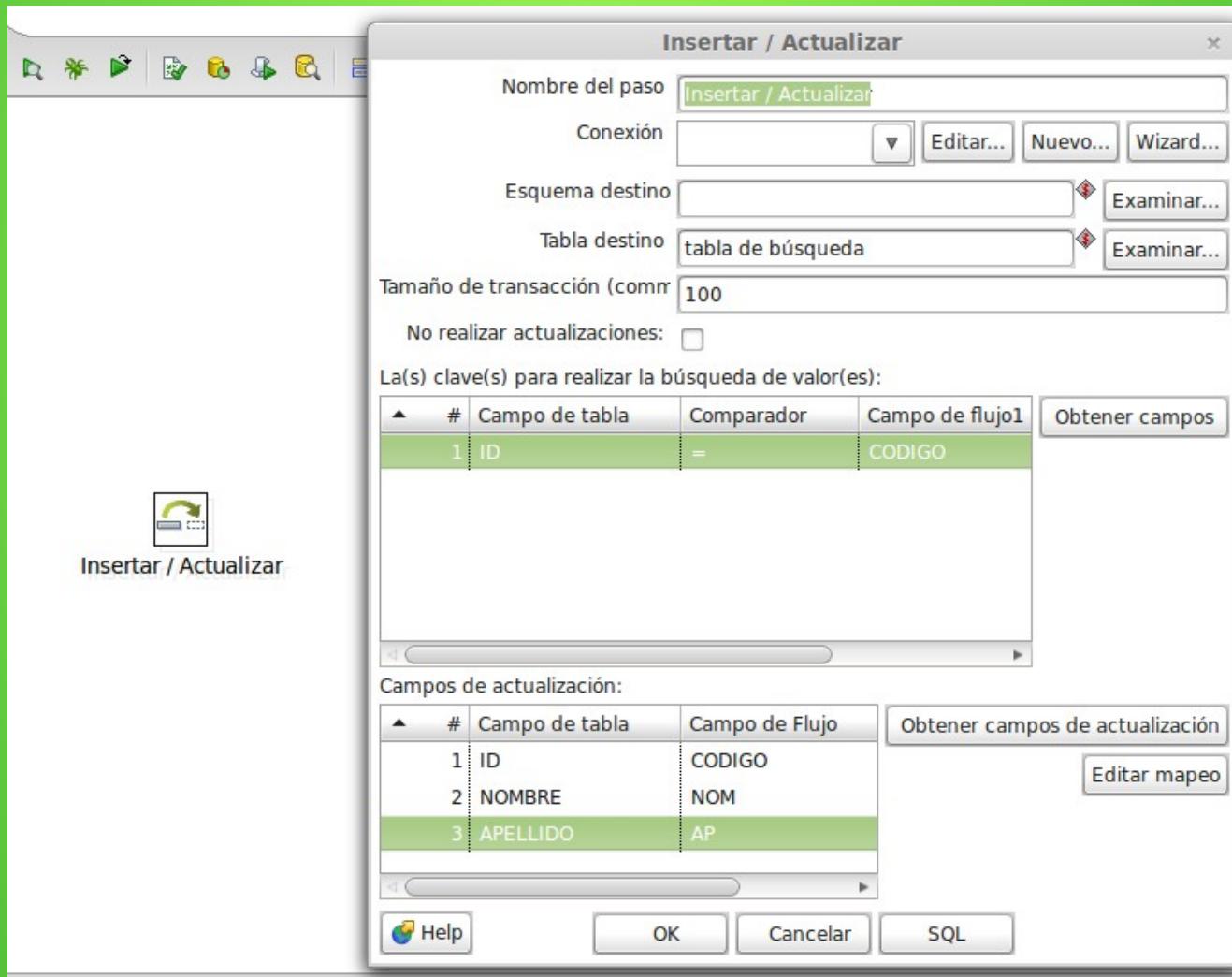
# PDI Pasos de SALIDA

## SALIDA A TABLA



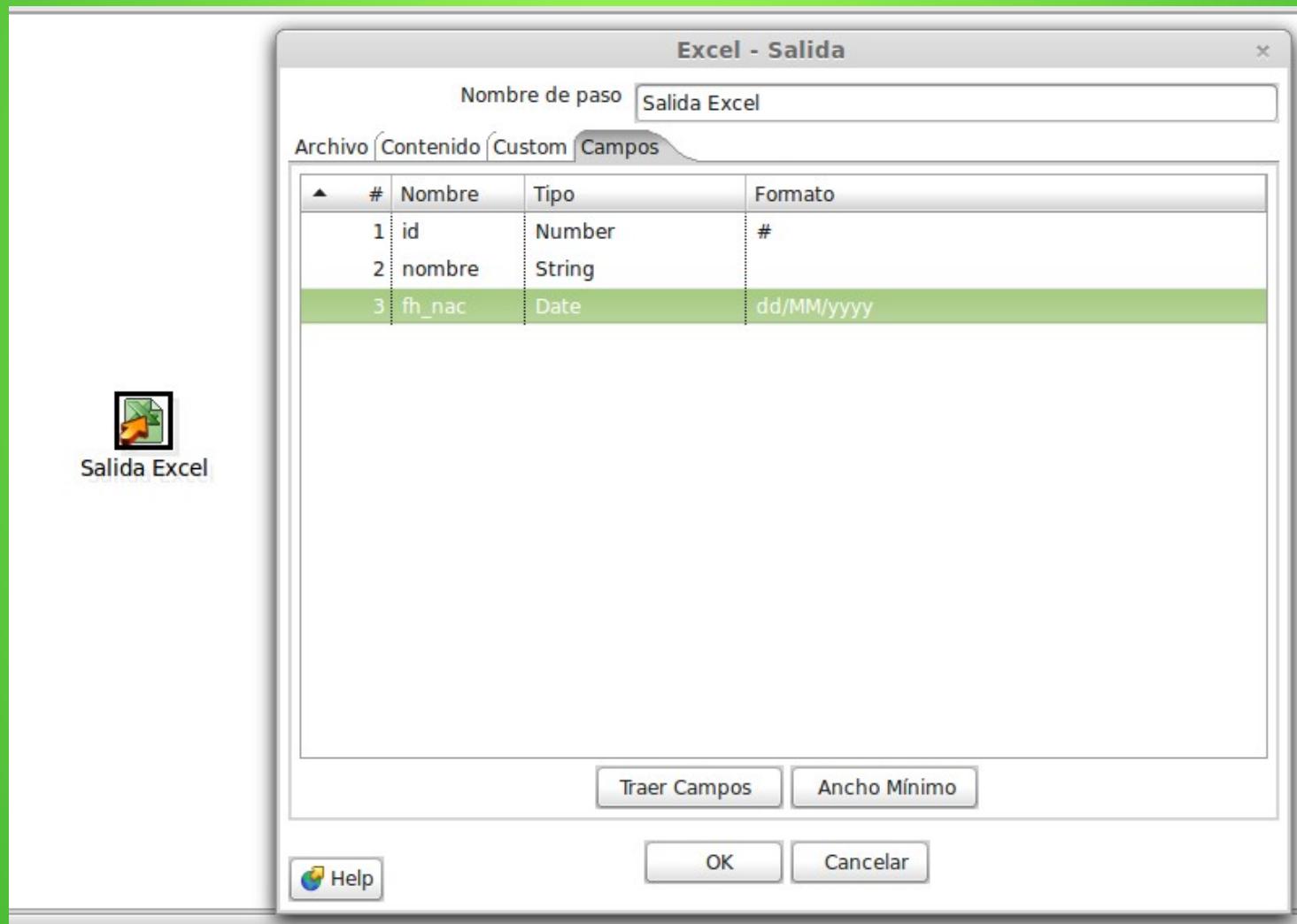
# PDI Pasos de SALIDA

## INSERTAR / ACTUALIZAR



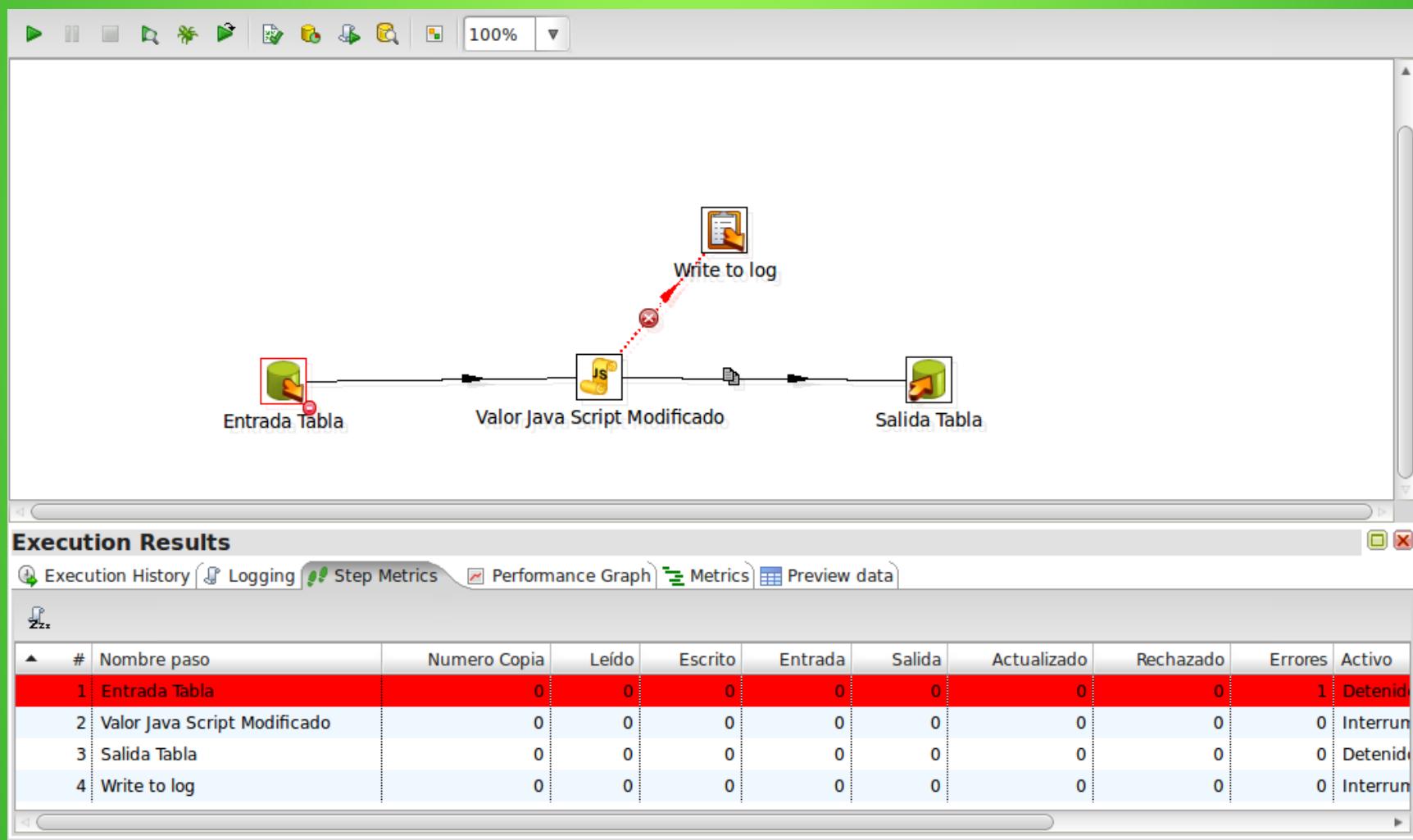
# PDI Pasos de SALIDA

## SALIDA A EXCEL



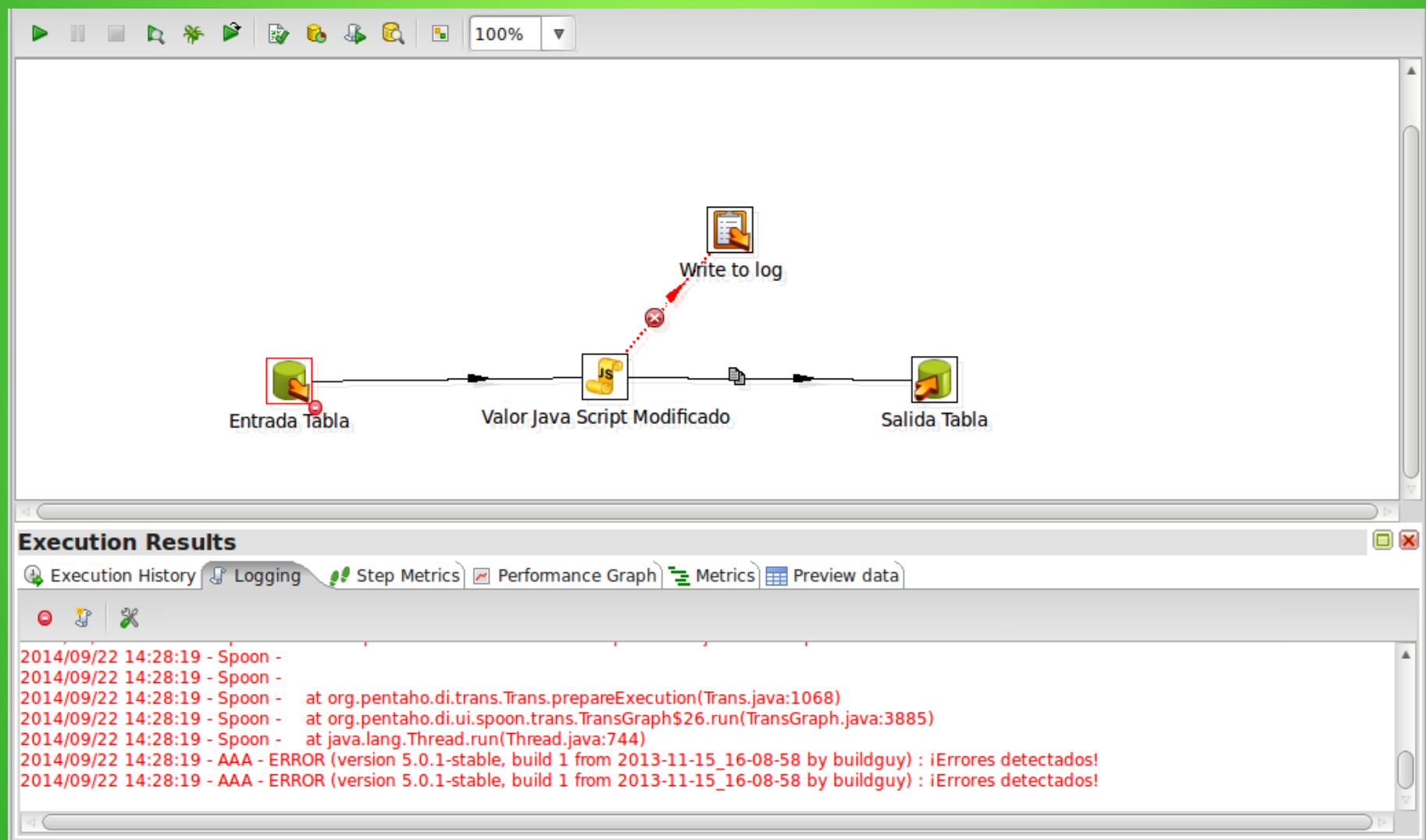
# PDI EJECUCION

# entrada - salida - leido - escrito



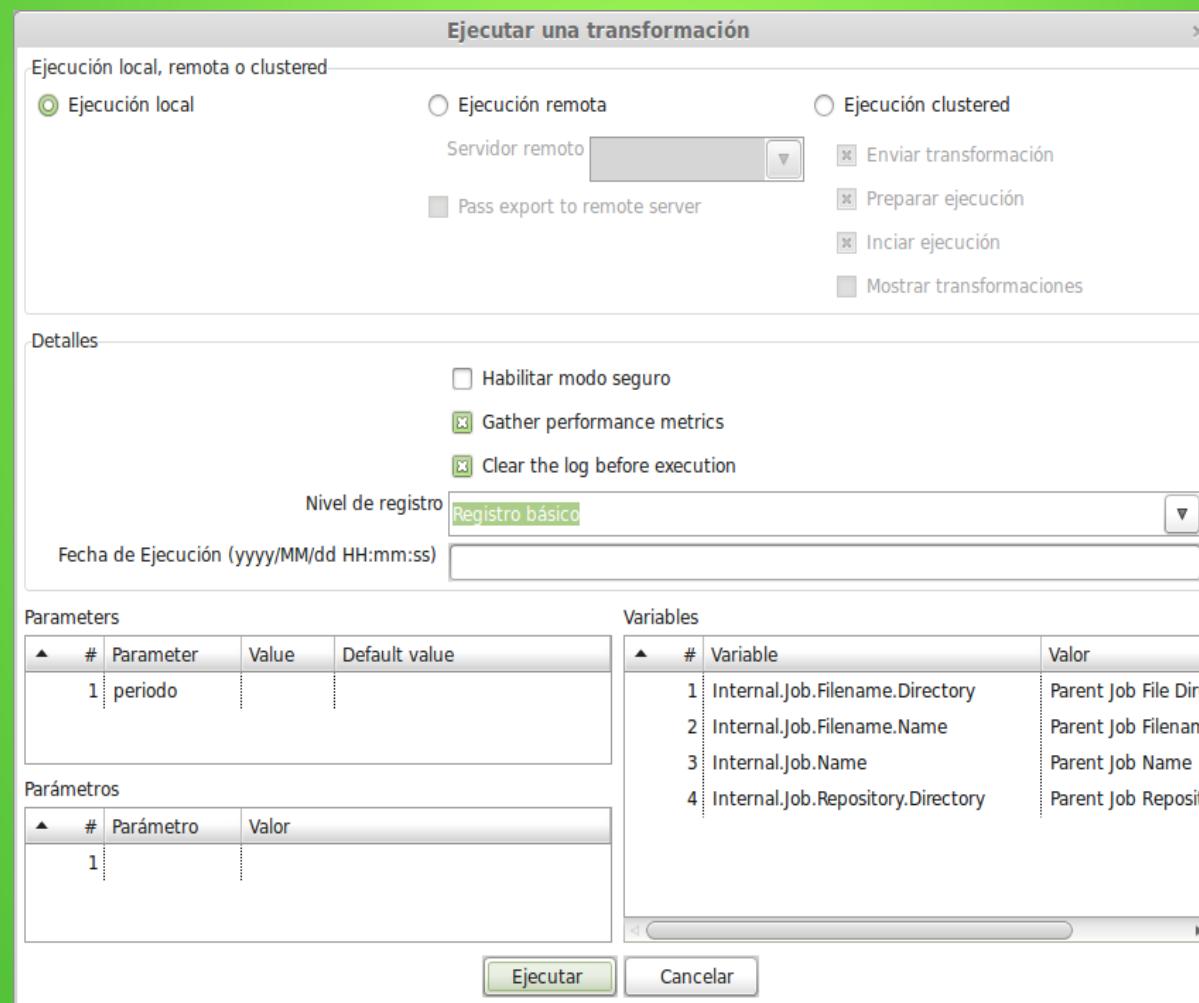
# PDI EJECUCION

## LOG DE ERRORES



# PDI EJECUCION

## PARAMETROS



# PDI EJECUCION

## PARAMETROS

The diagram illustrates the configuration of a 'Tabla - Entrada' (Table Input) step in PDI. On the left, two nodes are shown: 'Asignar valor a variables' (Assign value to variables) and 'Entrada Tabla' (Table Input). An arrow points from the 'Asignar valor a variables' node to the 'Entrada Tabla' node, indicating that variable values are being assigned to the input step.

**Tabla - Entrada (Table Input) Configuration:**

- Nombre paso:** Entrada Tabla
- Conexión:** (Connection dropdown menu with options: Editar..., Nuevo..., Wizard...)
- SQL:** SELECT \* FROM data WHERE periodo=?
- Line 1 Column 34:** (Text area showing the SQL query line and column information)
- Activar conversión perezosa:**
- ¿Reemplazar variables en script?**
- Insertar datos del paso:** Obtener Variables
- ¿Ejecutar para cada fila?**
- Limitar tamaño:** 0

At the bottom of the dialog are buttons for Help, OK, Previsualizar (Preview), and Cancelar (Cancel).

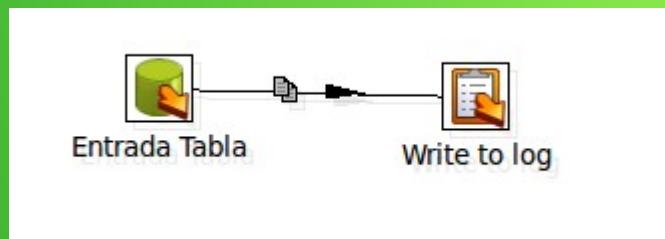
# PDI EJECUCION

## DEBUG

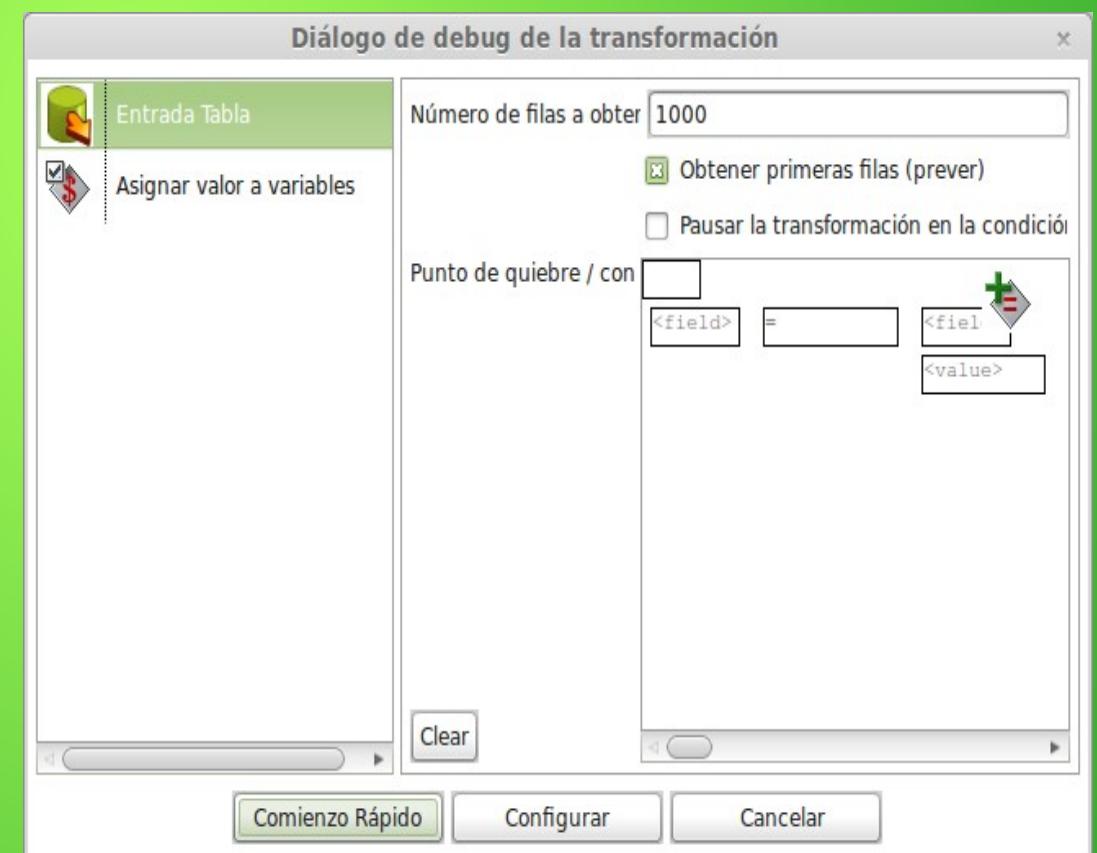
VER PASOS DE ENTRADA

VER PASOS DE SALIDA

ESCRIBIR AL LOG



## PREVISUALIZAR

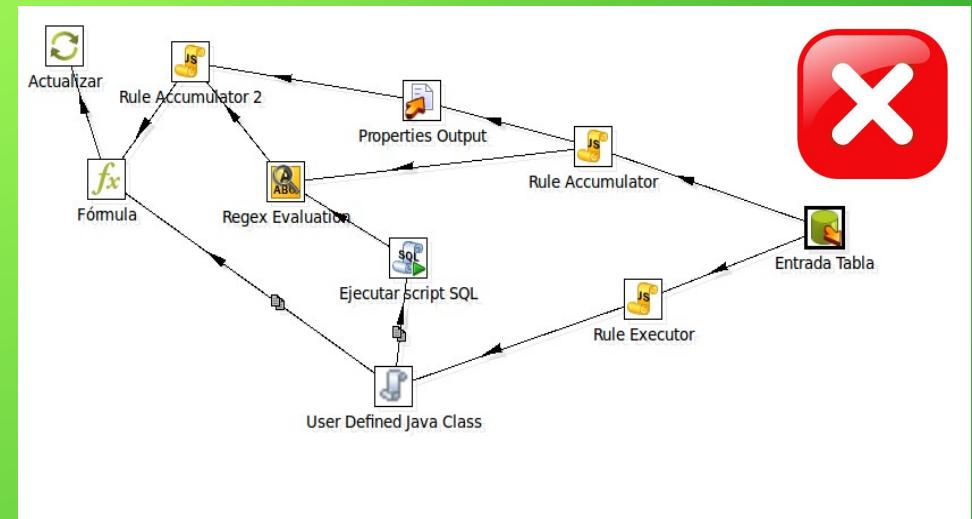
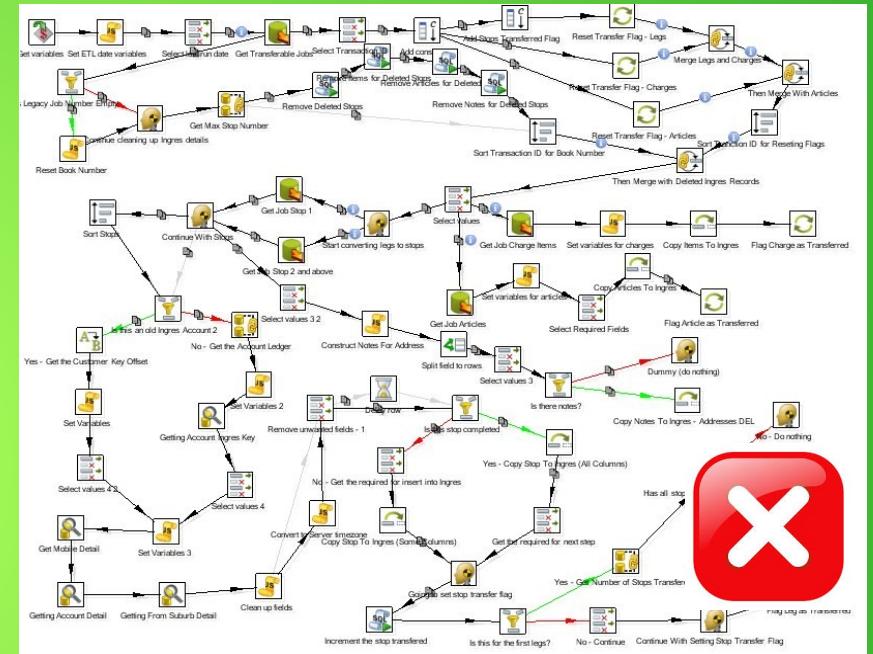
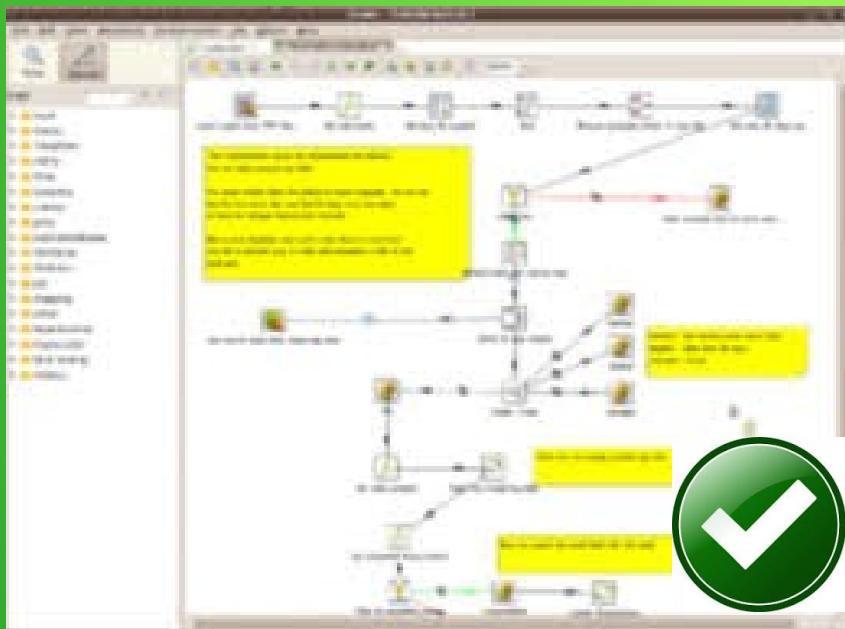
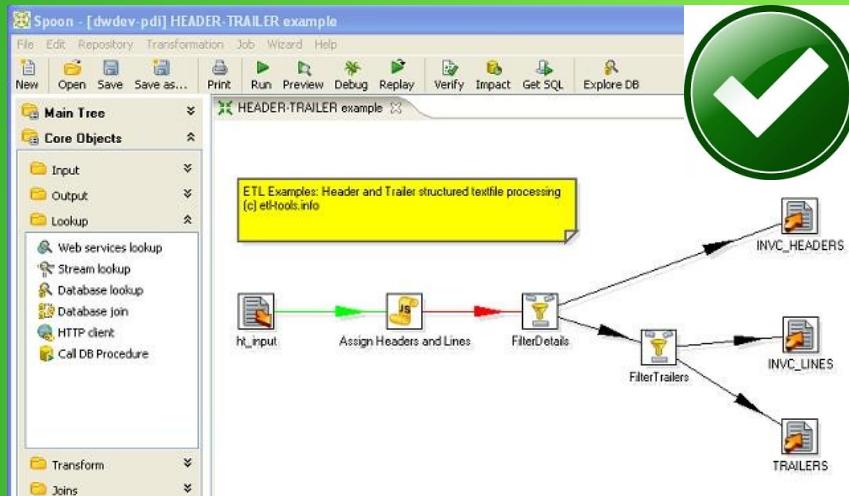


# PDI

## 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 homepage of the "Latest Pentaho Data Integration (aka Kettle) Documentation". At the top, there's a navigation bar with links to "Dashboard", "Pentaho Data Integration", and the current page. On the right, there are "Search", "View", "Edit", "Log In", and other user interface icons. Below the navigation, the title "Latest Pentaho Data Integration (aka Kettle) Documentation" is displayed next to the pentaho logo. A section for translated documentation includes links to "Spanish" (with a flag icon), "Korean" (with a flag icon), and "Russian". It also shows the "Current CI (Jenkins) build status (master branch)" which is marked as "Unstable". The main content area is divided into several sections: "Quick Start and Overview" (listing links to Home Page, Getting Started, FAQ, Case Studies, Recorded Demos, Screenshots, Articles, Tutorials, Concepts, Best Practices, and Solutions, and InfoCenter), "Documentation for (Database) Developers and Users", "Reference" (listing links to Spoon User Guide, Pentaho Data Integration Steps, Pentaho Data Integration Job Entries, Pan User Documentation, and Kitchen User Documentation), and "Community".

### 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

