Pentaho Data Integration

PDI

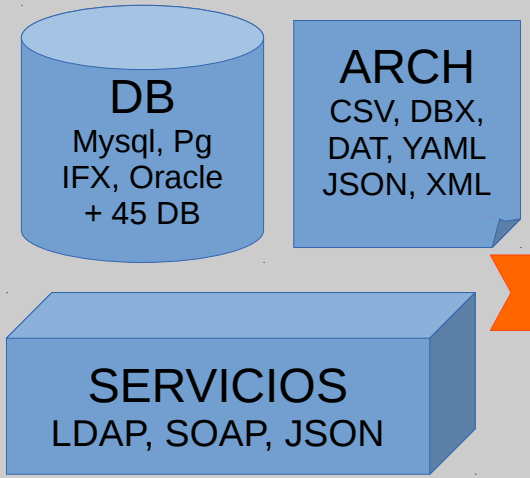
Herramienta de la solución PENTAHO
que se ocupa de los procesos ETL

Se utiliza para poblar DWH
Migrar Datos
Exportar vistas

ETL

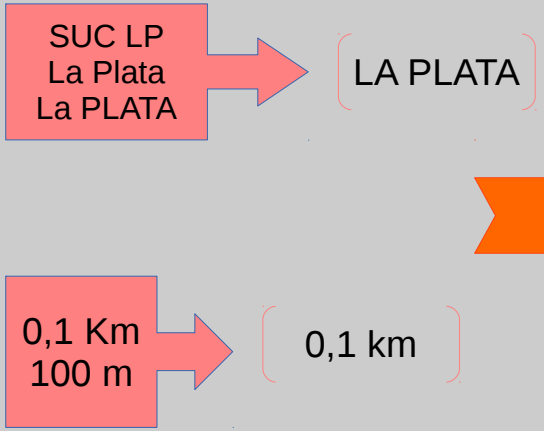
E
Extraction

Extrae datos de una o mas fuentes de datos de origen



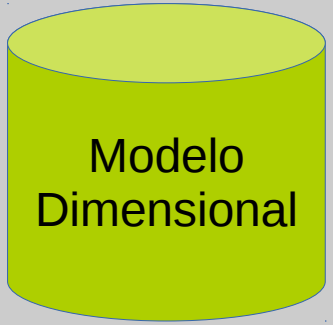
T
Transformation

Transforma los datos para que sean consistentes y fáciles de interpretar



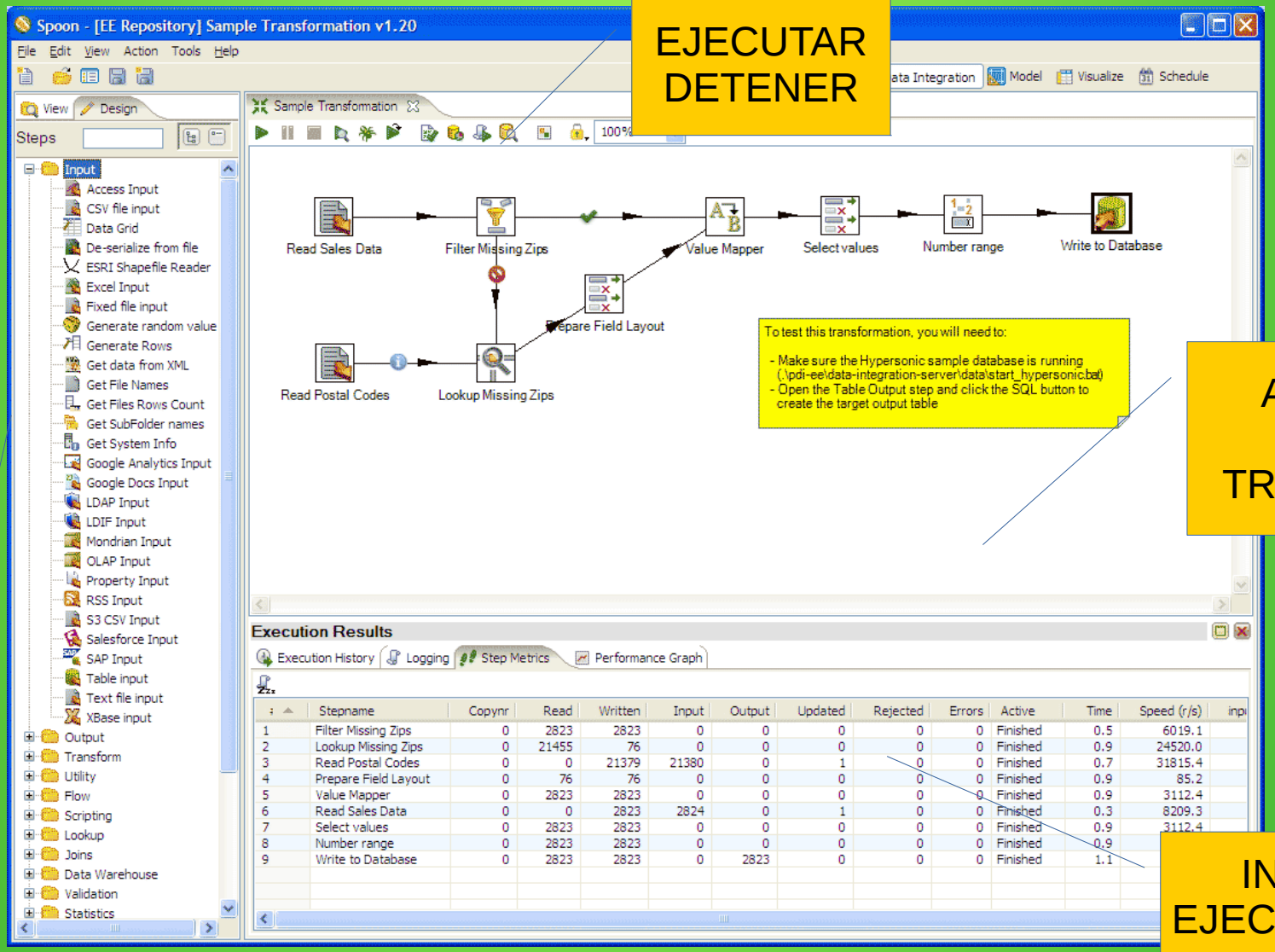
L
Load

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



PDI

JAVA - Multiplataforma



EJECUTAR
DETENER

PASOS

AREA
DE
TRABAJO

INFO
EJECUCION

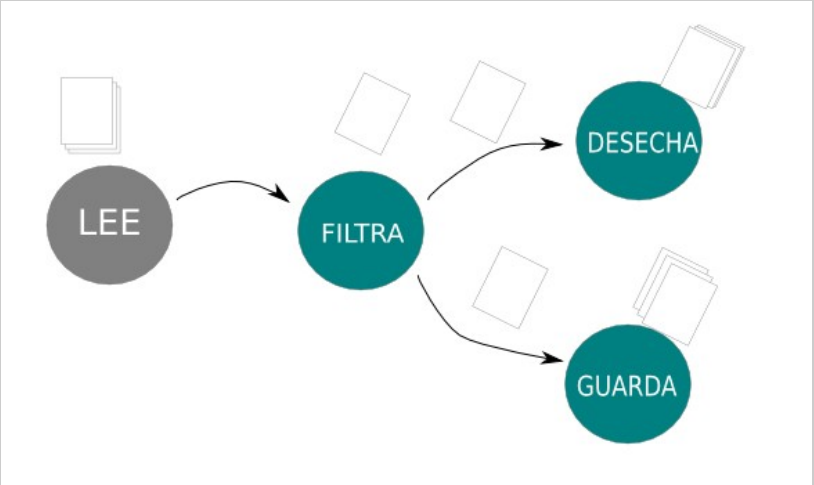
PDI

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

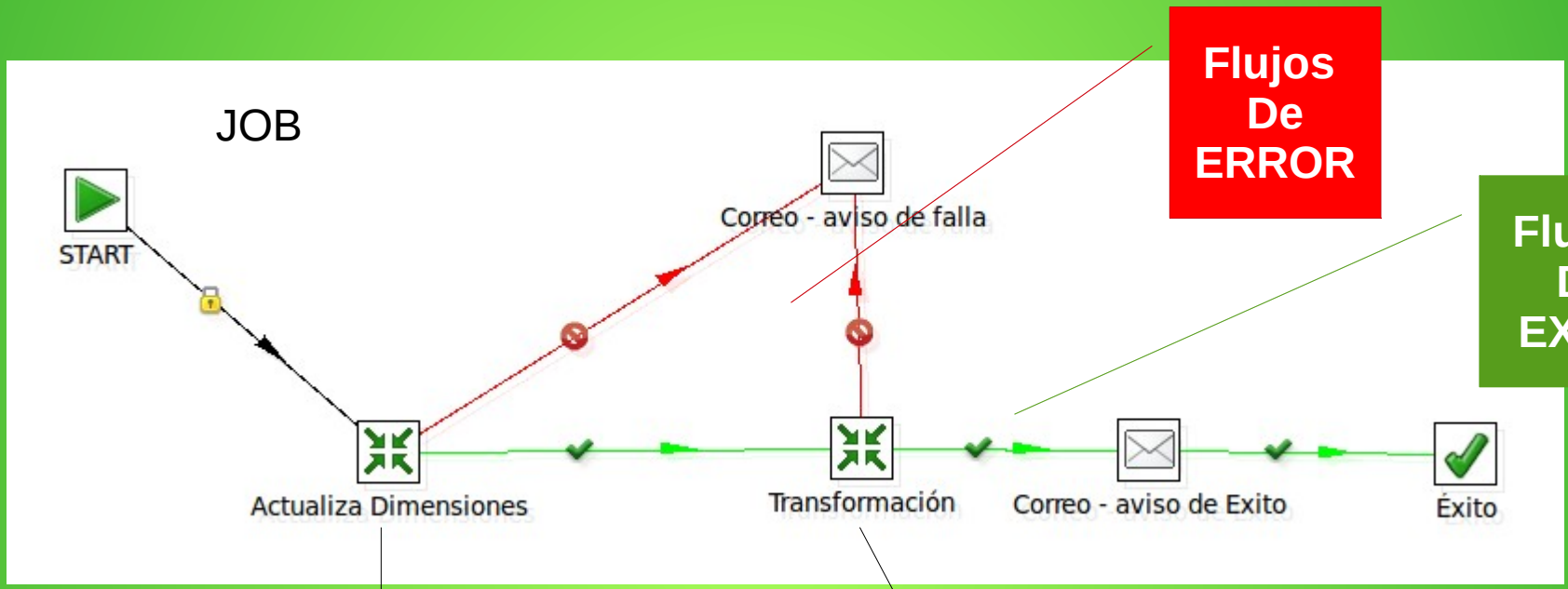


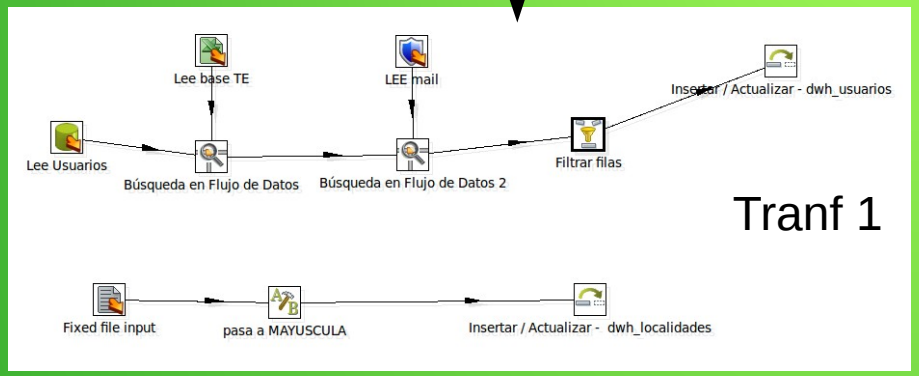
Trabajos
Kettle Job *.kjb

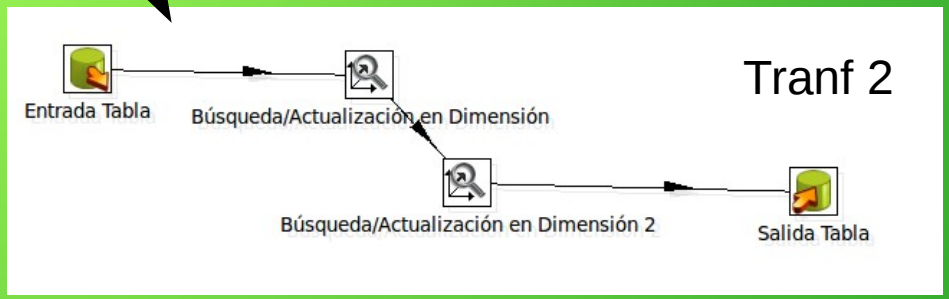
Se ejecuta de un paso a la vez.

Se utiliza para organizar varias transformaciones

PDI


JOB
START
Correo - aviso de falla
Flujos De ERROR
Flujos De EXITO
Actualiza Dimensiones
Transformación
Correo - aviso de Exito
Éxito


Lee base TE
LEE mail
Insertar / Actualizar - dwh_usuarios
Lee Usuarios
Búsqueda en Flujo de Datos
Búsqueda en Flujo de Datos 2
Filtrar filas
Tranf 1
Fixed file input
pasa a MAYUSCULA
Insertar / Actualizar - dwh_localidades

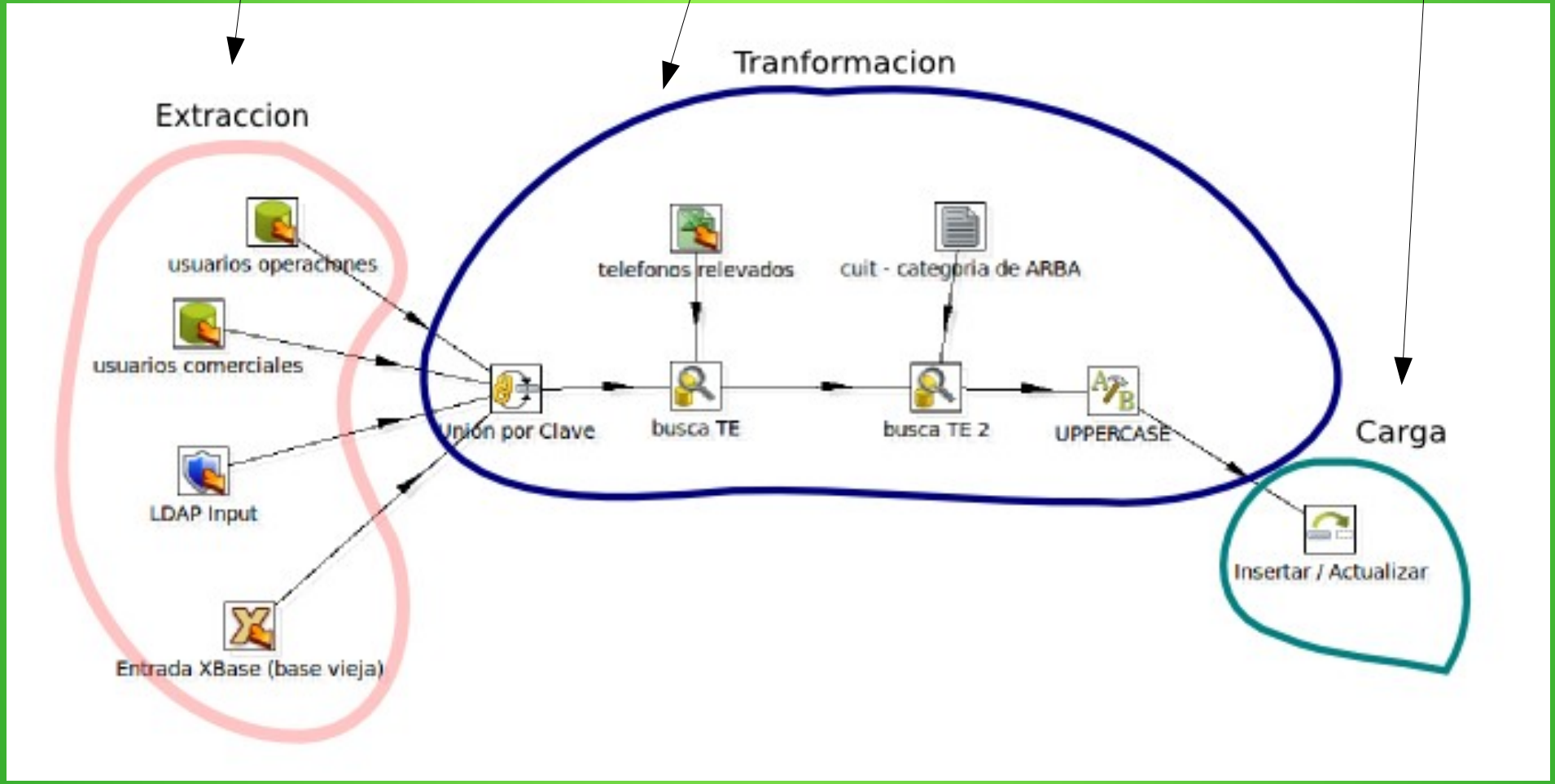

Entrada Tabla
Búsqueda/Actualización en Dimensión
Tranf 2
Búsqueda/Actualización en Dimensión 2
Salida Tabla

PDI

Pasos de ENTRADA
Generan filas

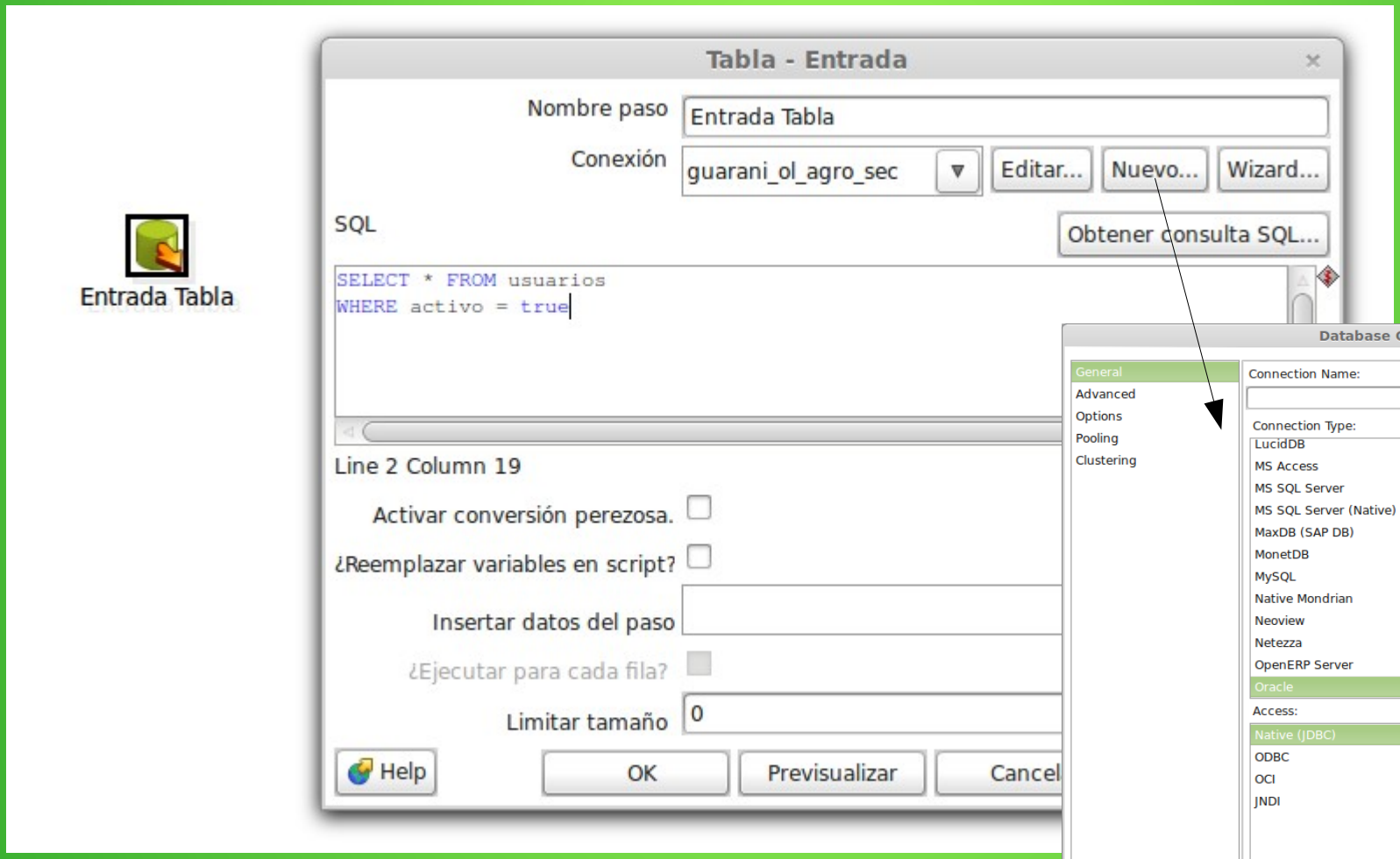
Pasos de TRANSF
Modifican filas

Pasos de SALIDA
GUARDAN filas



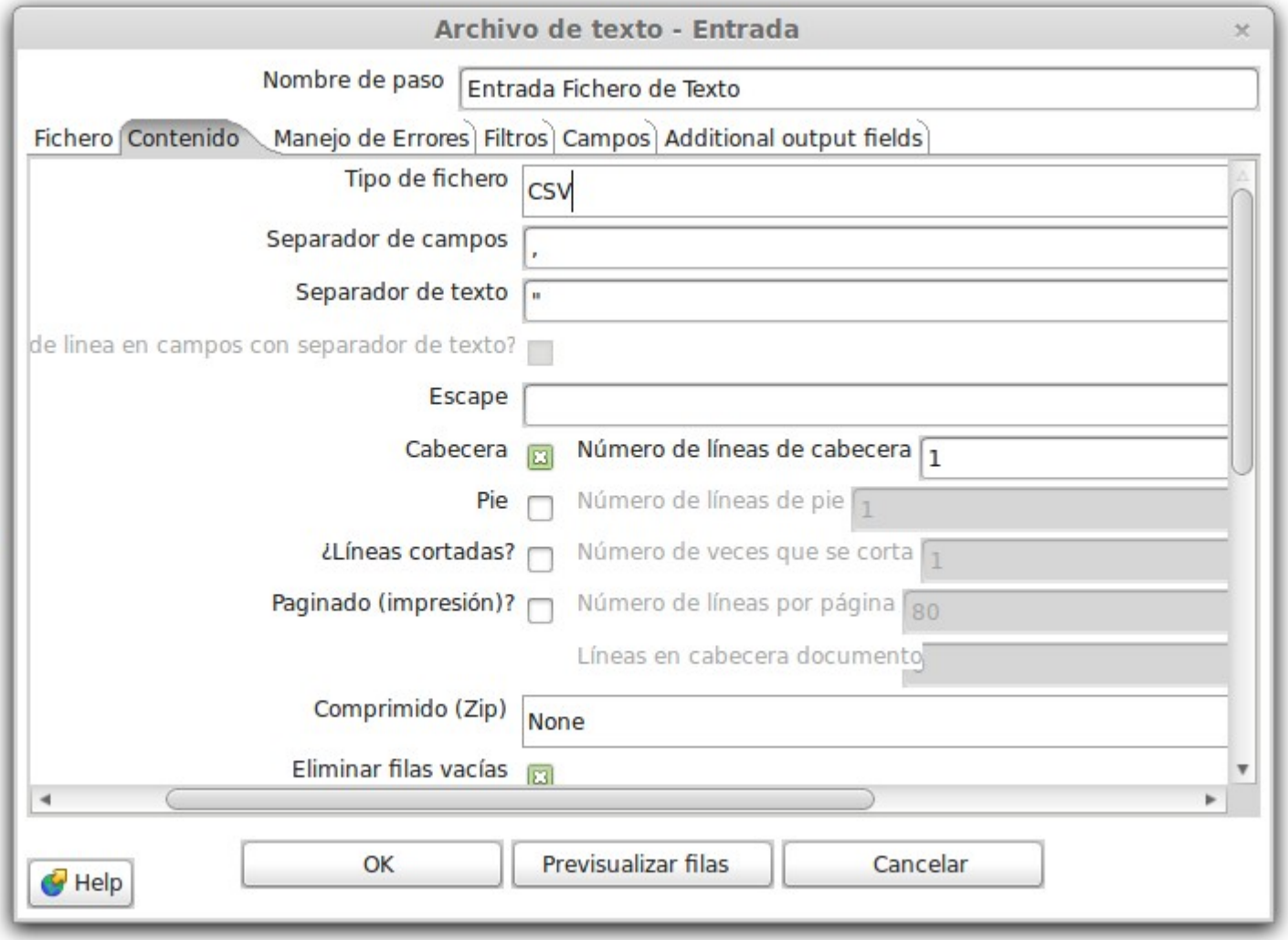
PDI Pasos de ENTRADA

Consulta a Base de Datos



PDI Pasos de ENTRADA

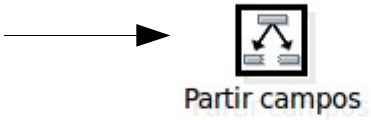
Lectura de Archivo



PDI Pasos de TRANSF

PARTIR CAMPOS

razon_soc
ROJAS, Ariel
MORA, Rodrigo

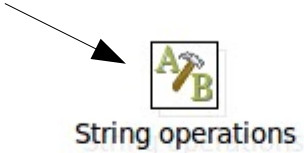


razon_soc	**Nombre**	**Apellido**
ROJAS, Ariel	**Ariel**	**ROJAS**
MORA, Rodrigo	**Rodrigo**	**MORA**

PDI Pasos de TRANSF

OPERACIONES SOBRE STRINGS

Localidad	cod
La plata	29
Buenos aires	0021

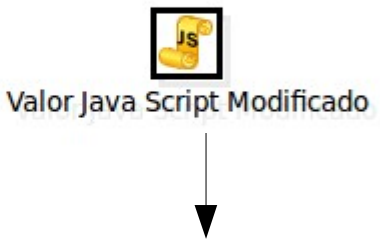


Localidad	cod
La Plata	00029
Buenos Aires	00021

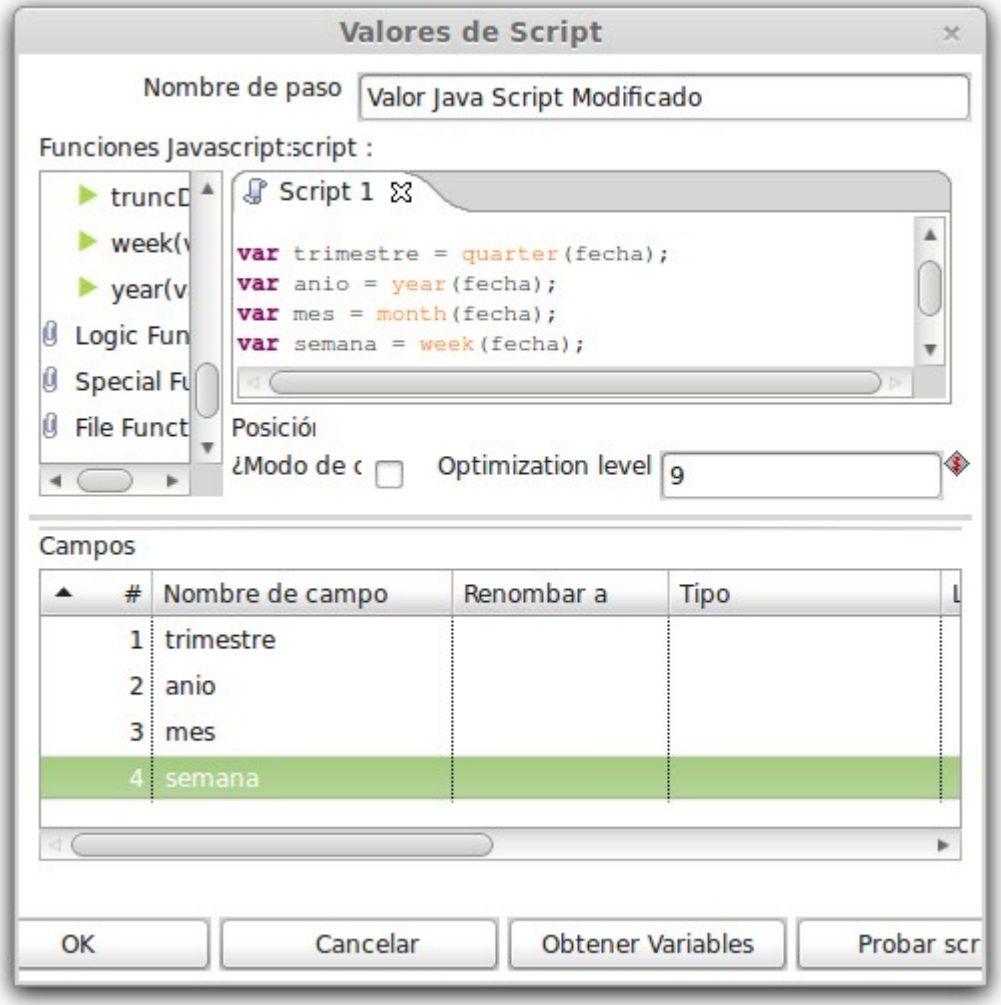
PDI Pasos de TRANSF

SCRIPTING

fecha
1-1-2014
2-1-2014



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



PDI Pasos de TRANSF

UNION POR CLAVE

legajo	nombre
1900	Teofilo Gutierrez
1901	Matías Kranevitter

legajo

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

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

Algunos pasos, como este, requieren la entrada ordenada

PDI Pasos de TRANSF

BUSQUEDA EN FLUJO DE DATOS

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

Busca por mal

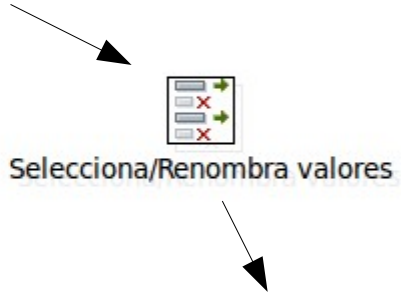
mail	asistencias
tg@gmail.com	12
mk@gmail.com	22

legajo	asistencias
1900	12
1901	22

PDI Pasos de TRANSF

SELECCIONA / RENOMBRA

c1	c2	c3	c4	c5	c6

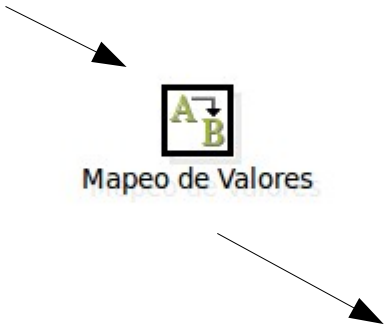


c5	c3	c1	c2

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

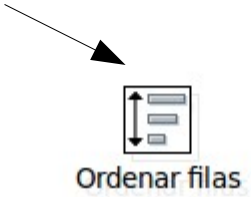


Genero	estudio	trabaja
Masculino	Primario	Trabaja
Femenino	Primario	No Trabaja
Masculino	Secundario	Trabaja
Masculino	Secundario	Trabaja
Femenino	Universitario	No Trabaja

PDI Pasos de TRANSF

ORDENA FILAS

dni	nombre
23456789	Marquezi, Jose
21234321	Rodriguez, Jorge



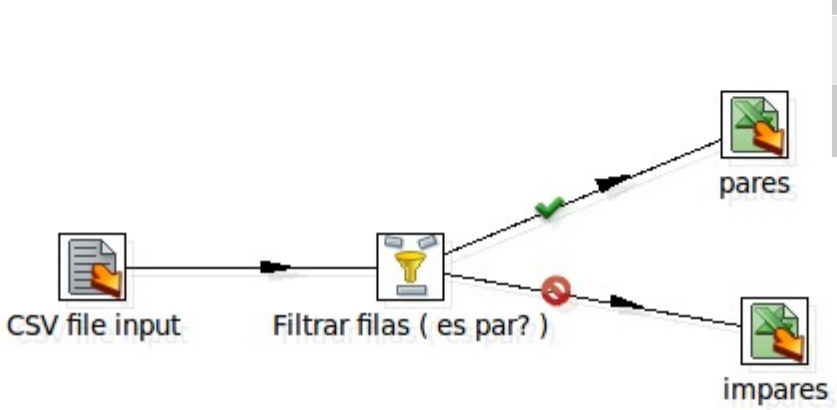
Este paso requiere obtener
TODAS las filas antes de continuar

dni	nombre
21234321	Rodriguez, Jorge
23456789	Marquez, Jose

PDI Pasos de TRANSF

FLUJO : FILTRAR FILAS

num
1
2
3
4
5
6
7

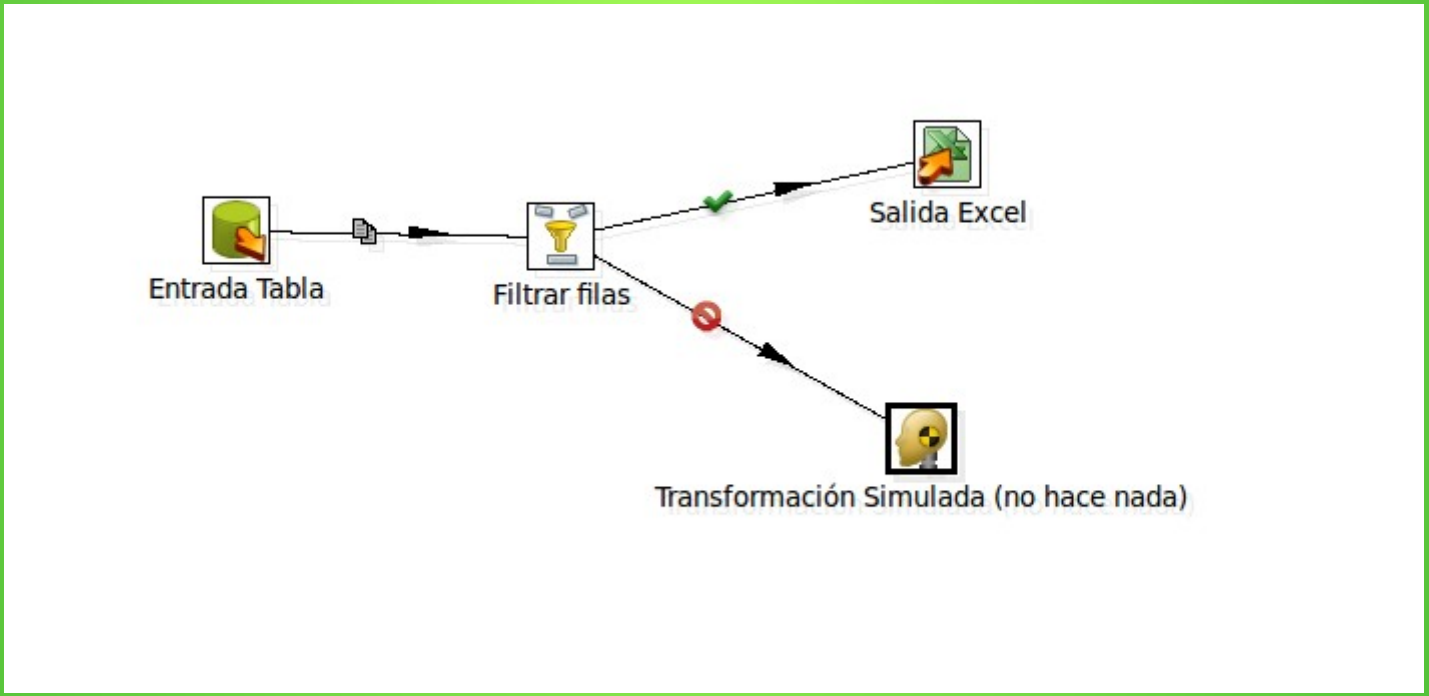


num
2
4
6

num
1
3
5
7

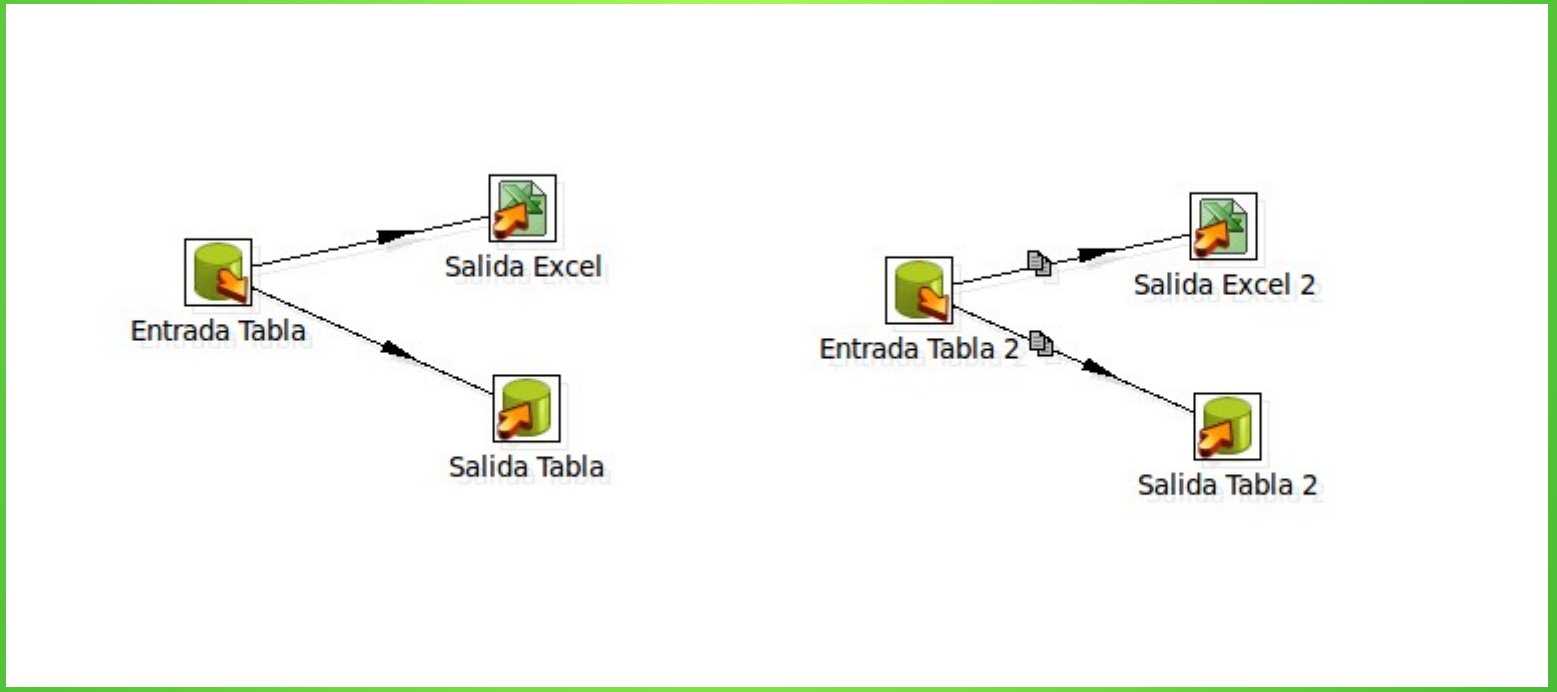
PDI Pasos de TRANSF

FLUJO : PASO NO HACE NADA



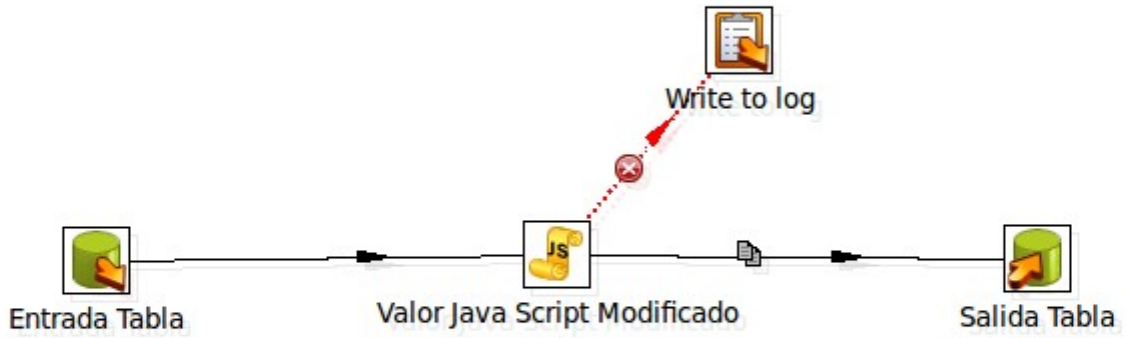
PDI Pasos de TRANSF

FLUJO : COPIA vs DISTRIBUYE



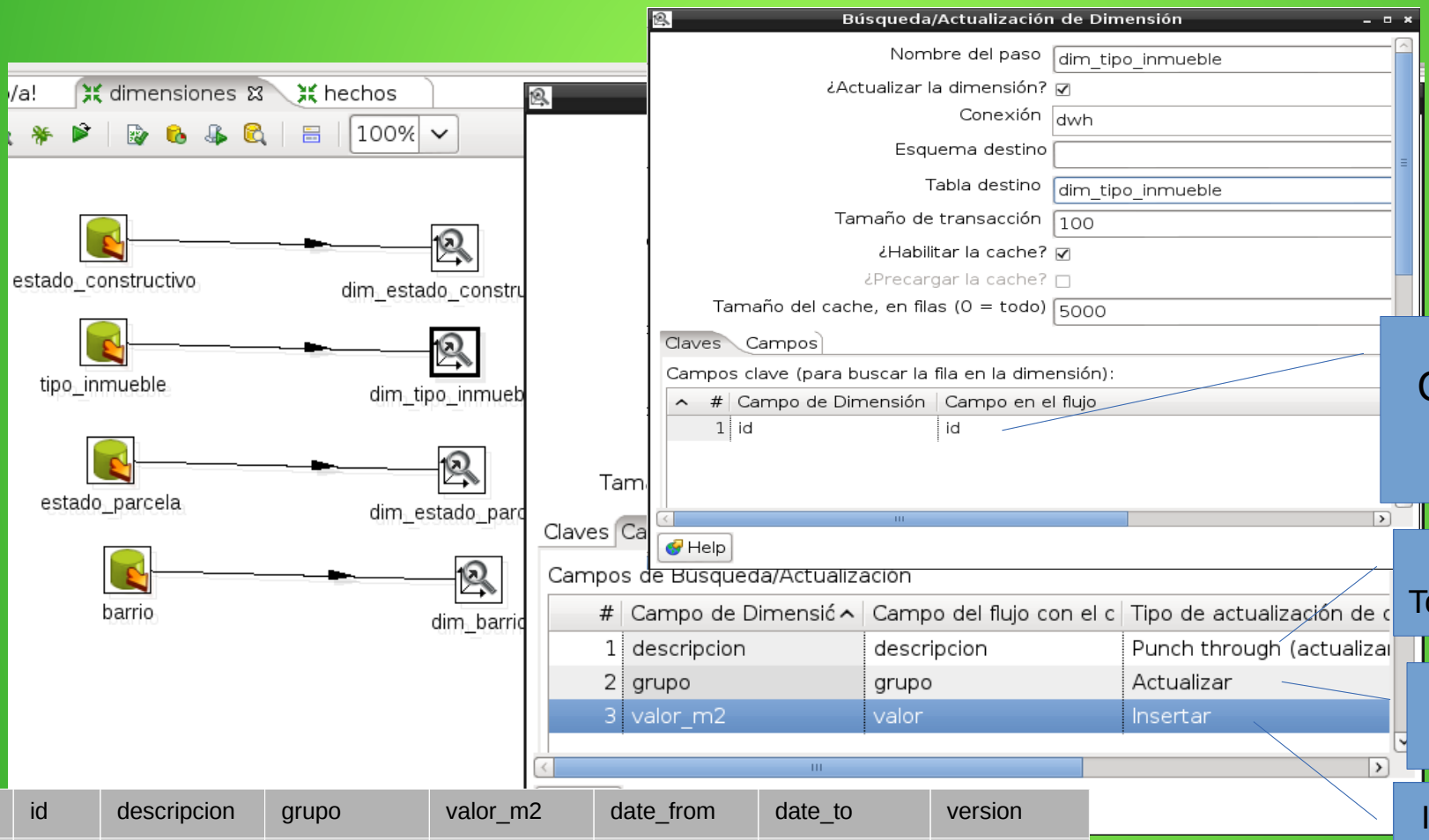
PDI Pasos de TRANSF

FLUJO : MANEJO DE ERRORES



PDI Pasos de TRANSF

DWH – Actualizar Dimension



Campo de búsqueda

Actualiza Todas las versiones

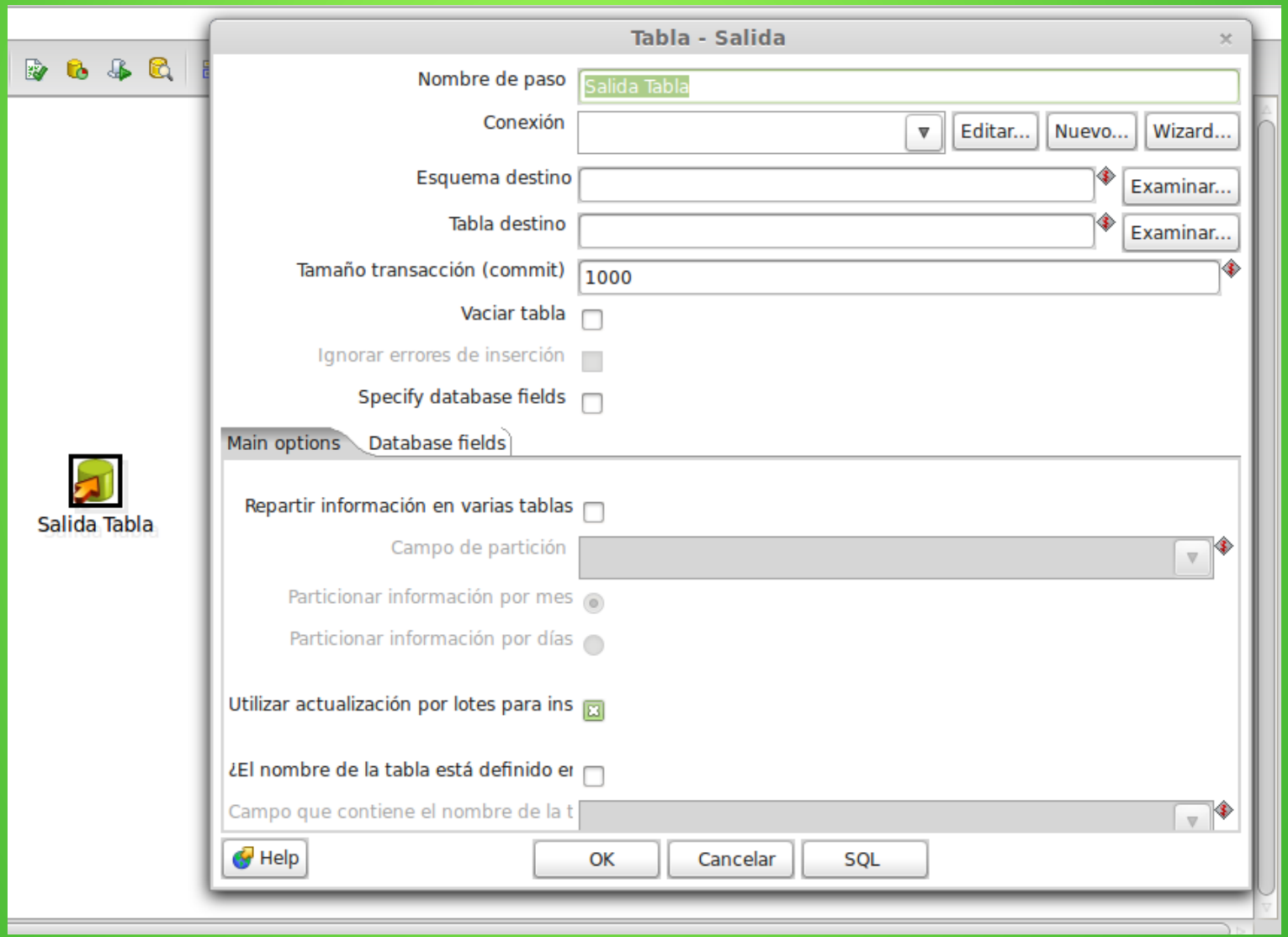
Actualiza la Ultima versión

Inserta una nueva versión del registro

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

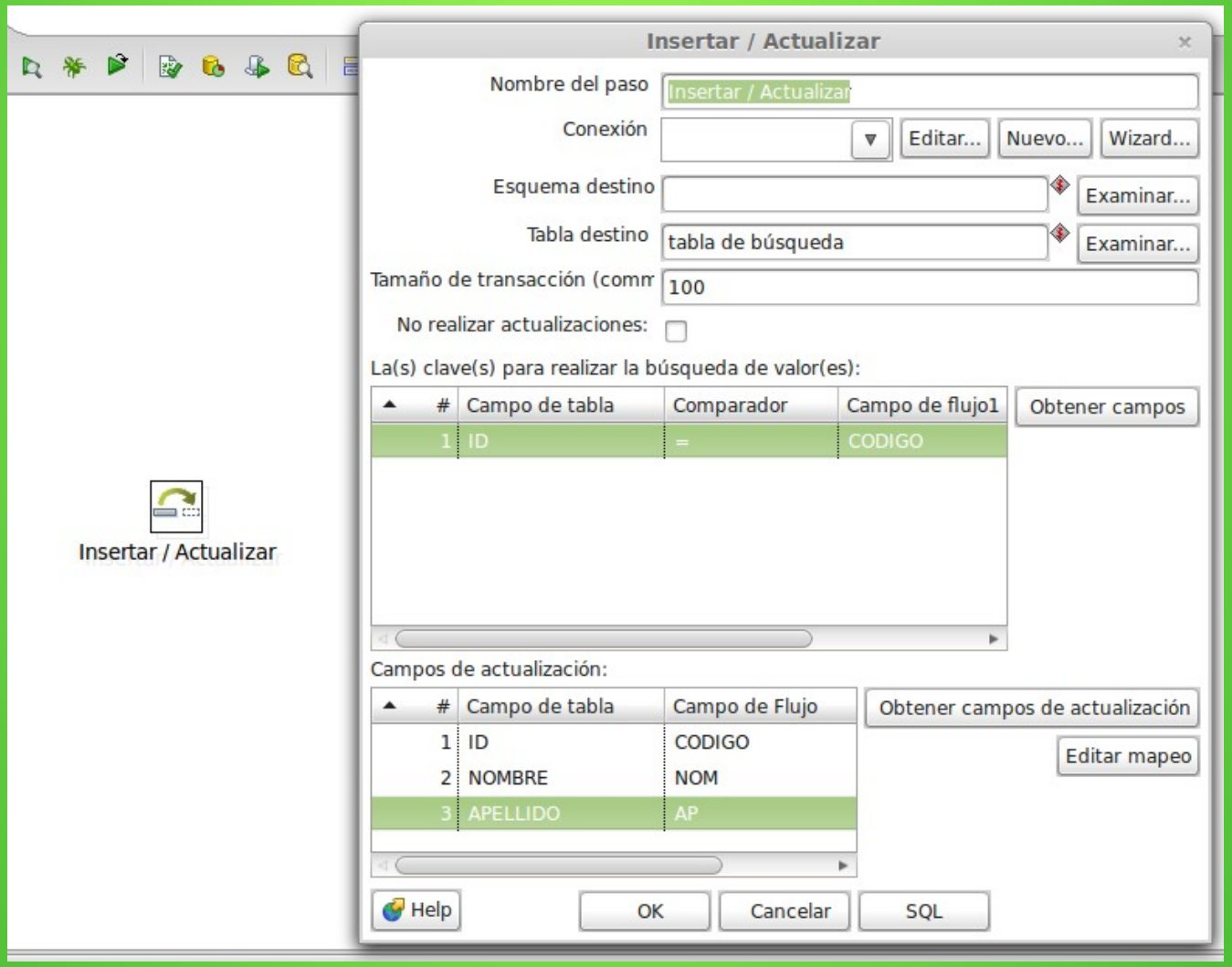
PDI Pasos de SALIDA

SALIDA A TABLA



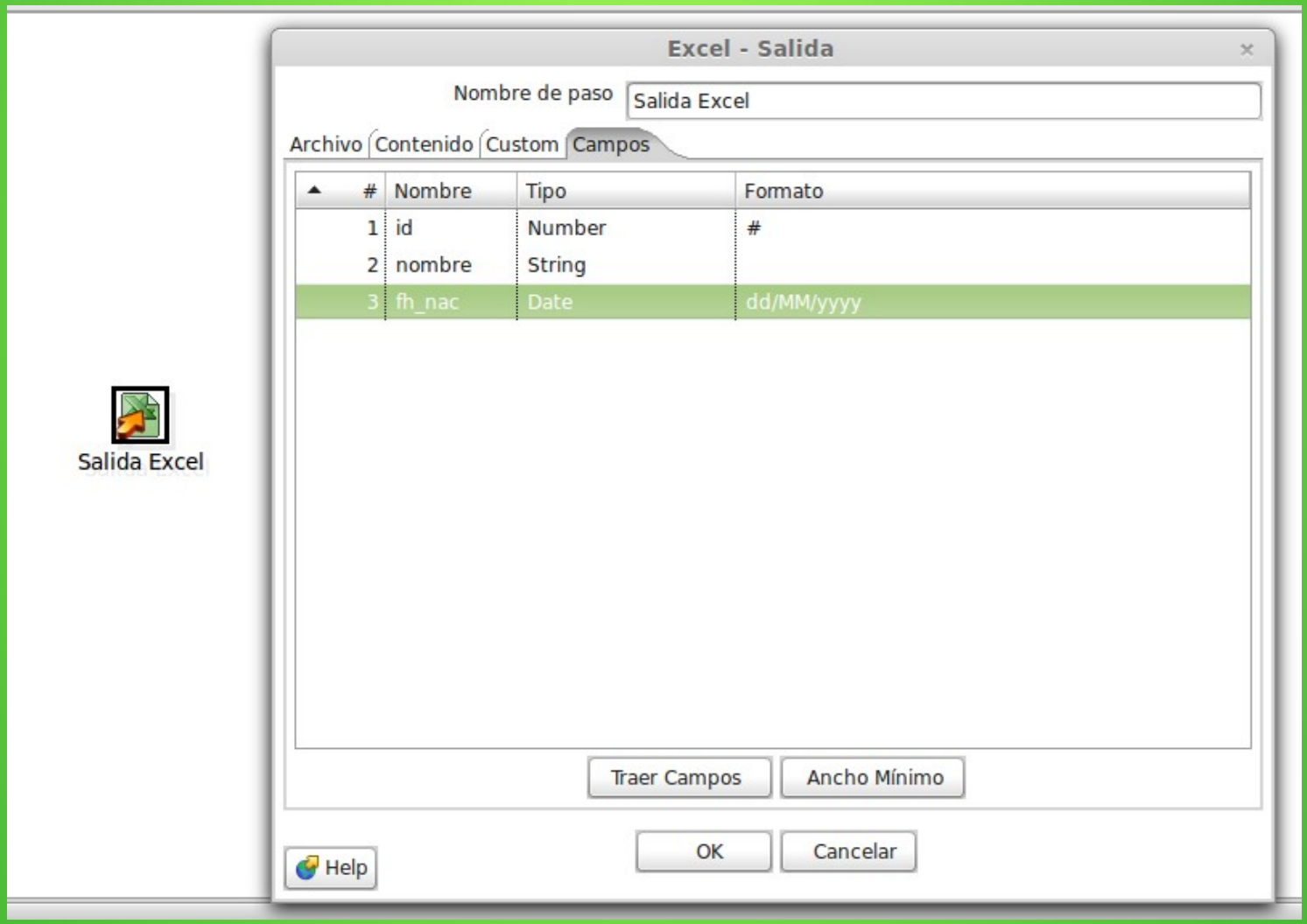
PDI Pasos de SALIDA

INSERTAR / ACTUALIZAR



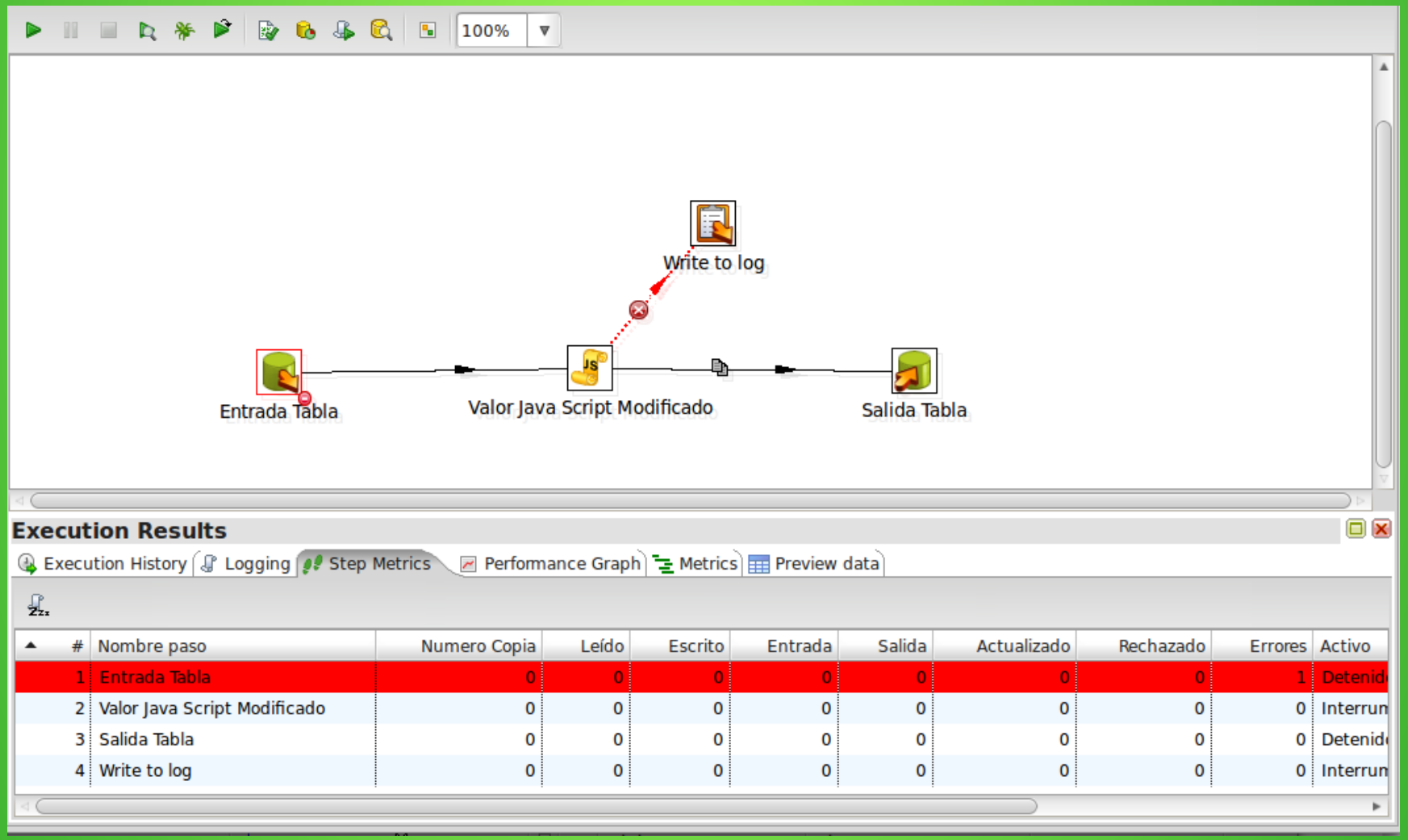
PDI Pasos de SALIDA

SALIDA A EXCEL



PDI EJECUCION

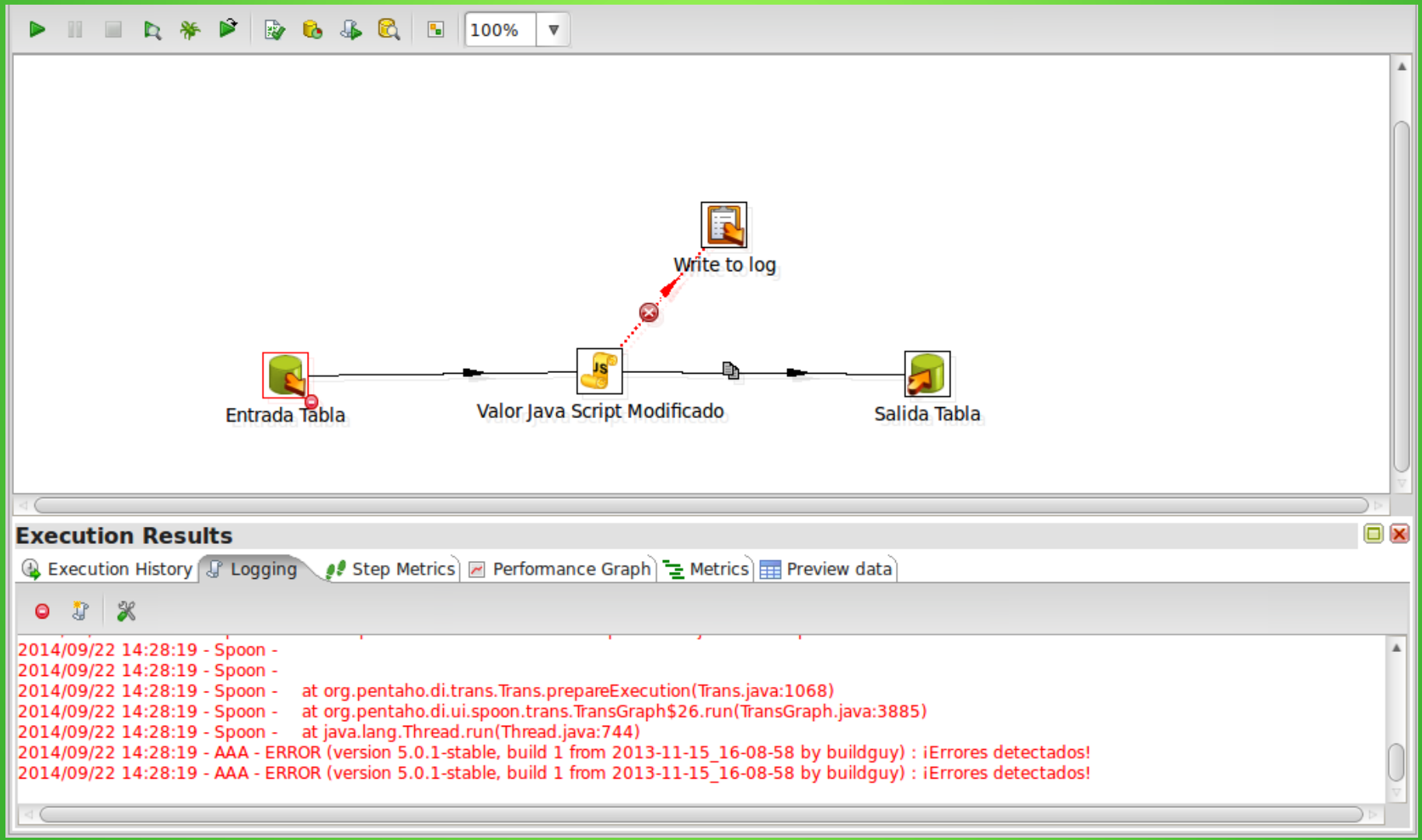
entrada - salida - leido - escrito



#	Nombre paso	Numero Copia	Leído	Escrito	Entrada	Salida	Actualizado	Rechazado	Errores	Activo
1	Entrada Tabla	0	0	0	0	0	0	0	1	Detenid
2	Valor Java Script Modificado	0	0	0	0	0	0	0	0	Interrun
3	Salida Tabla	0	0	0	0	0	0	0	0	Detenid
4	Write to log	0	0	0	0	0	0	0	0	Interrun

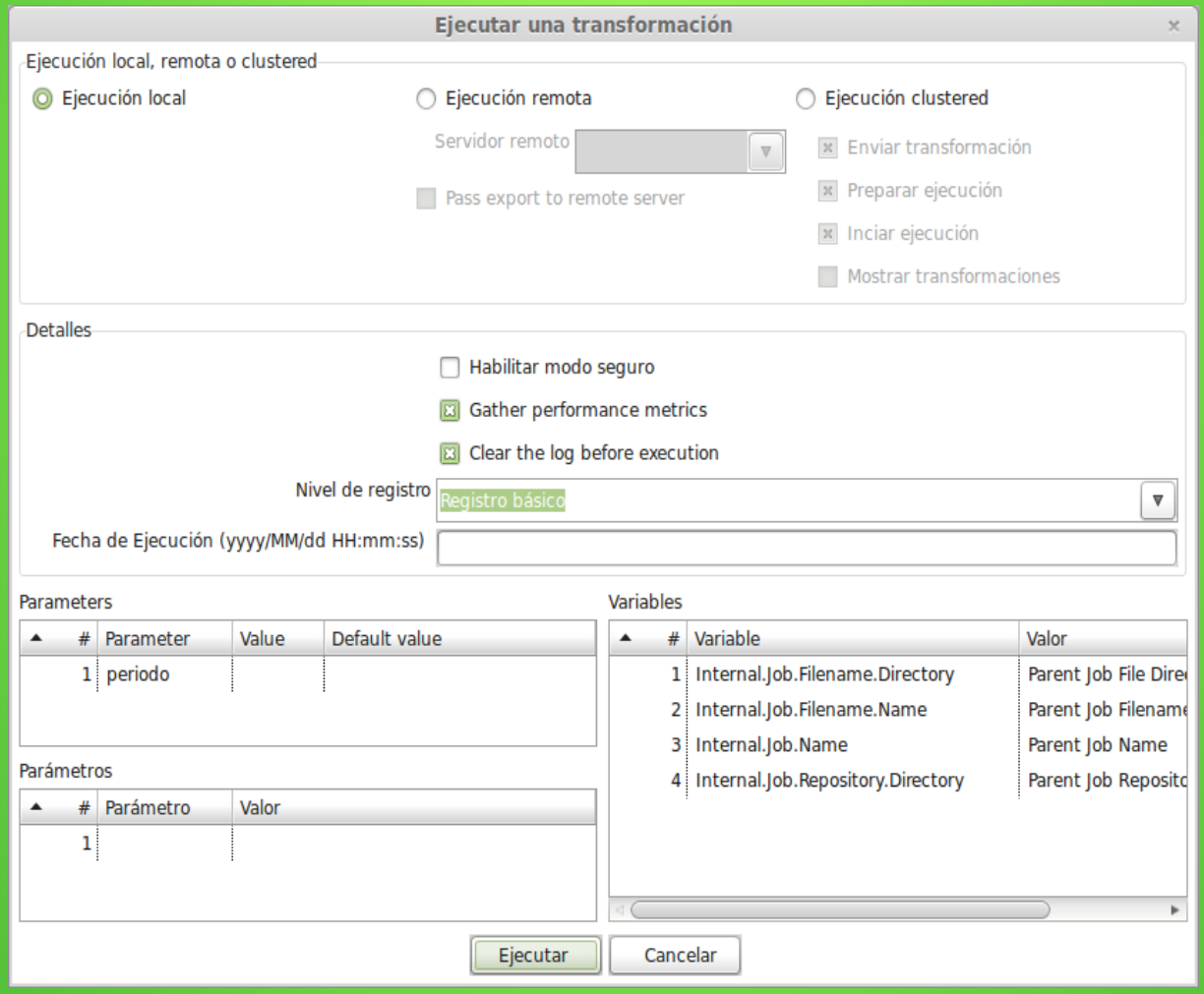
PDI EJECUCION

LOG DE ERRORES



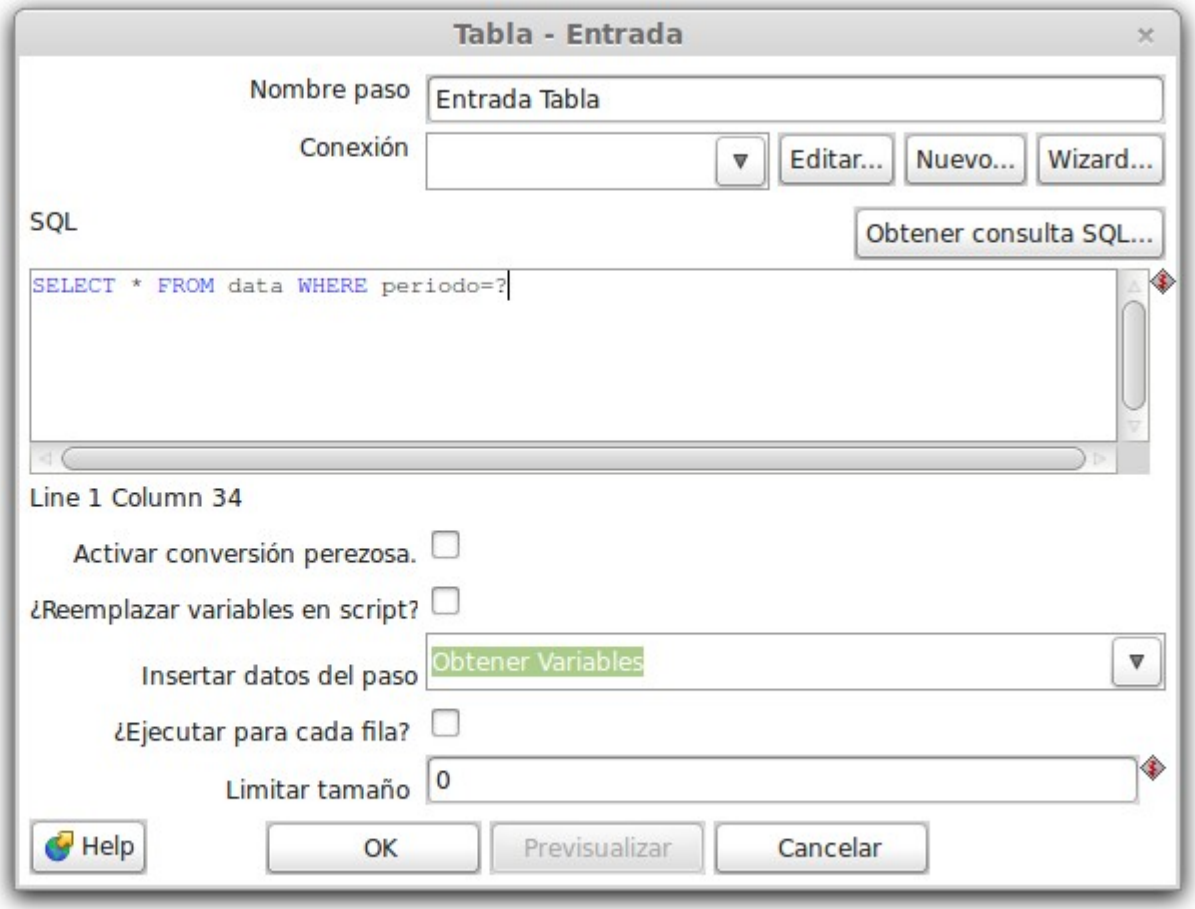
PDI EJECUCION

PARAMETROS



PDI EJECUCION

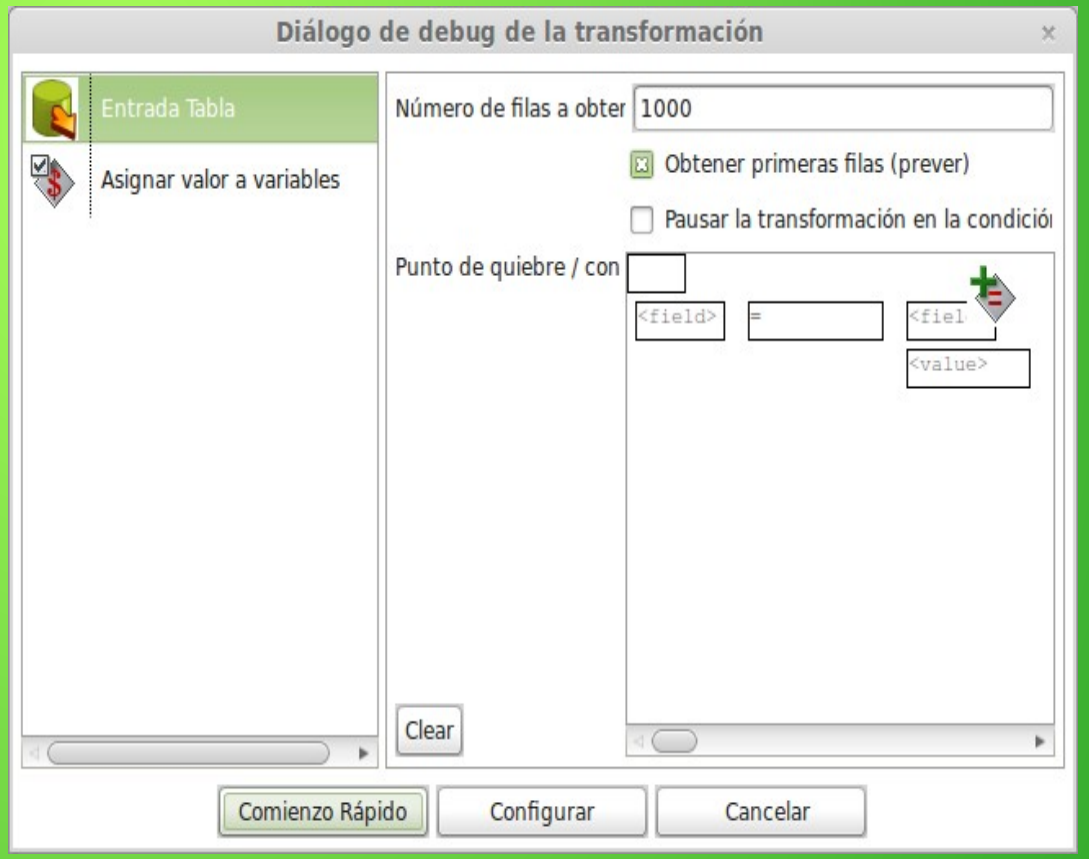
PARAMETROS



PDI EJECUCION

DEBUG

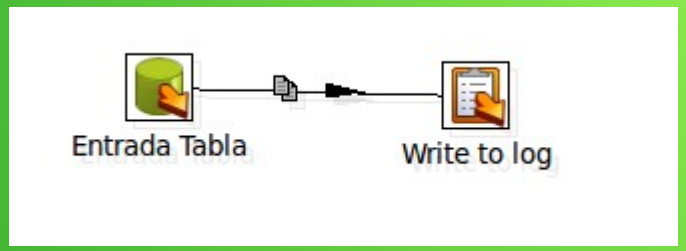
PREVISUALIZAR



VER PASOS DE ENTRADA

VER PASOS DE SALIDA

ESCRIBIR AL LOG

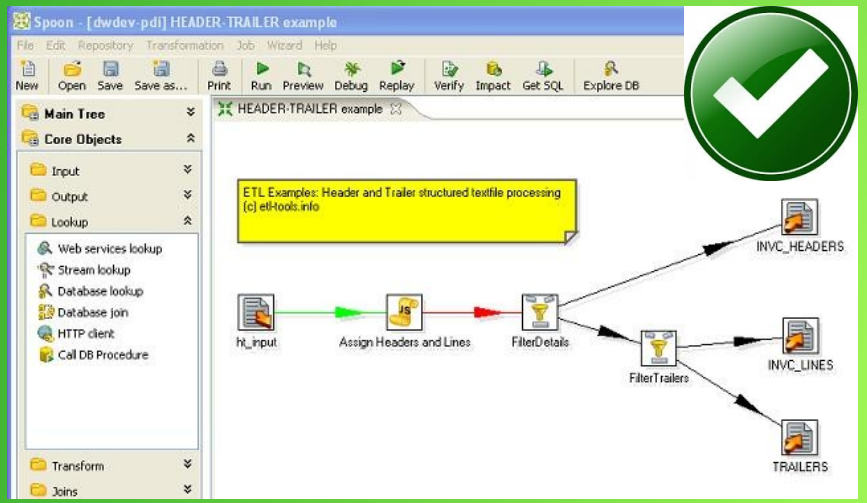


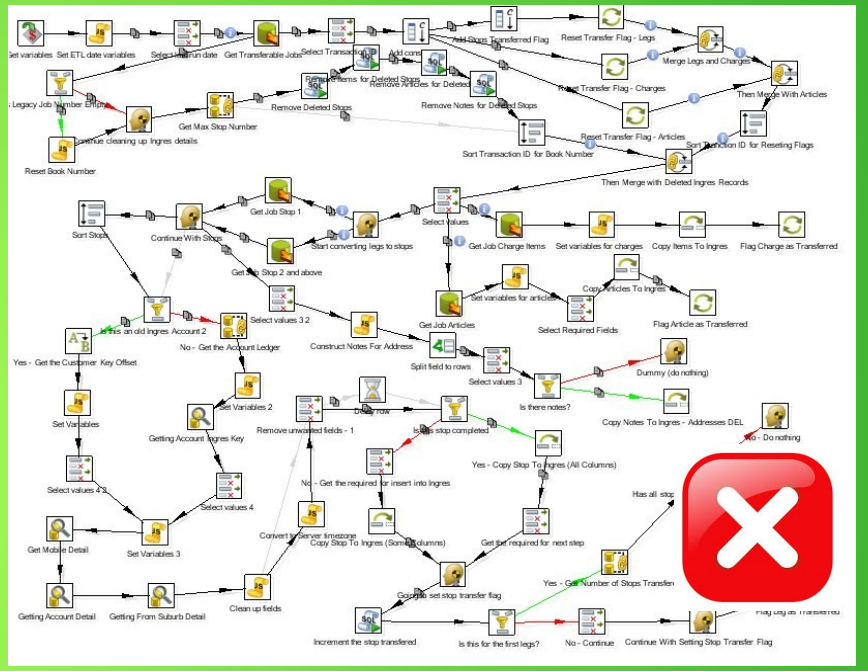
PDI

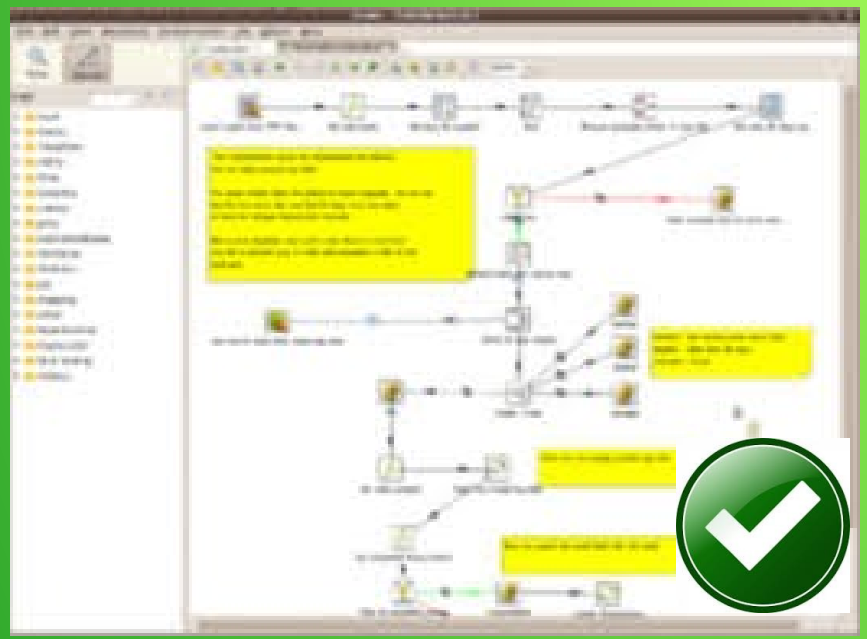
CONSEJOS UTILES

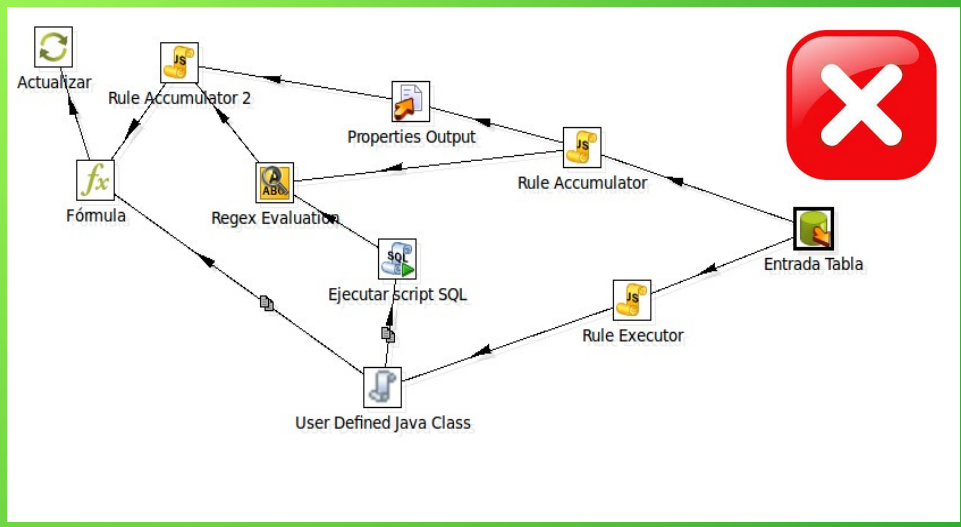
- Flujo de Izq → Der | Arriba → Abajo
- Utilizar nombres descriptivos para los pasos
- Documentar con Notas
- Mantener los archivos simples
- Separar en varias tranformaciones – jobs para simplificar
- Utilizar nombres vinculos relativos en lugar de absolutos
- Organizar los archivos para que se correspondan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

PDI - Ejemplos







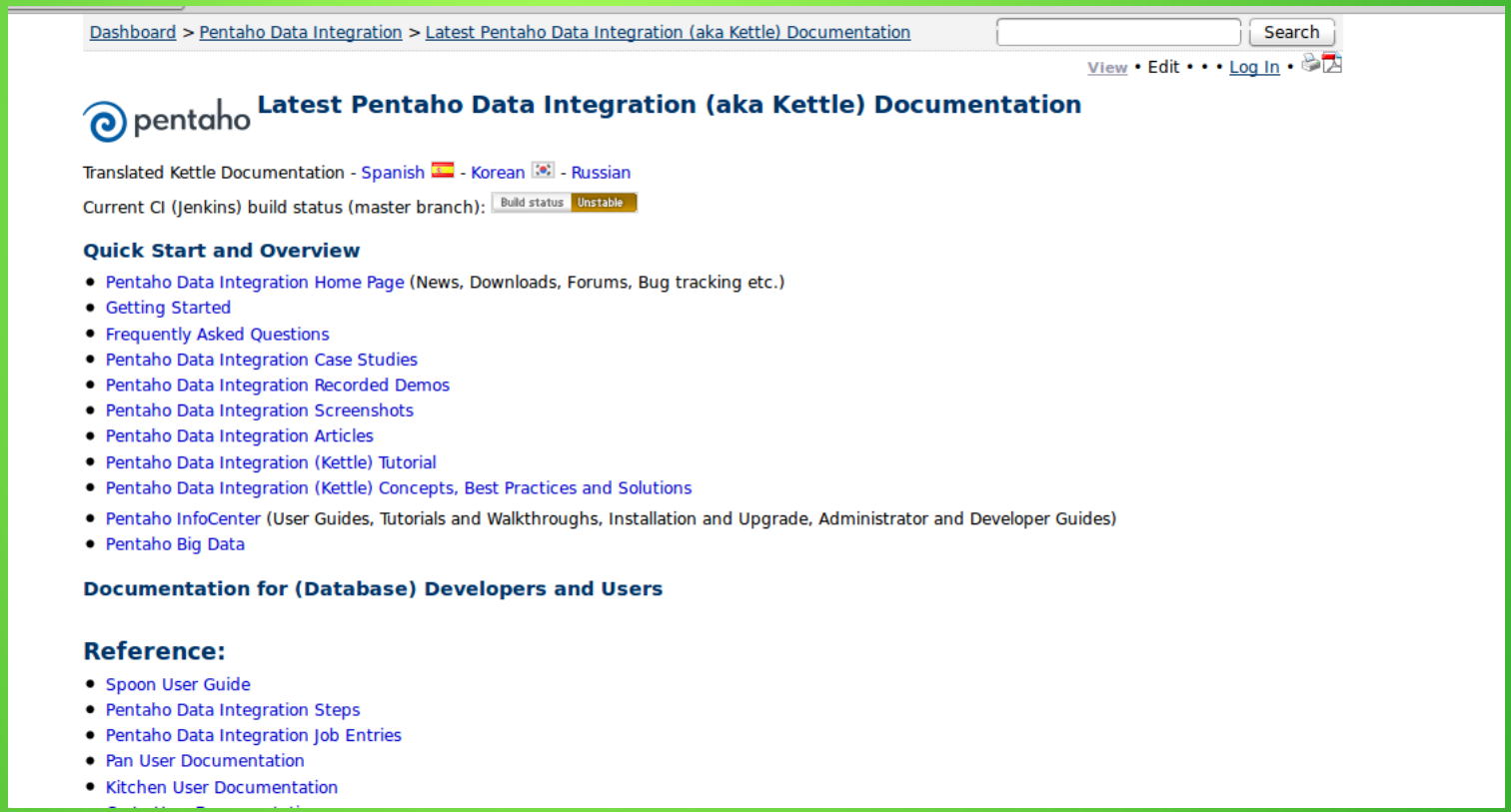


PDI

LINKS

DOCUNENTACION

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



DESCARGA

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

PDI

?

PDI

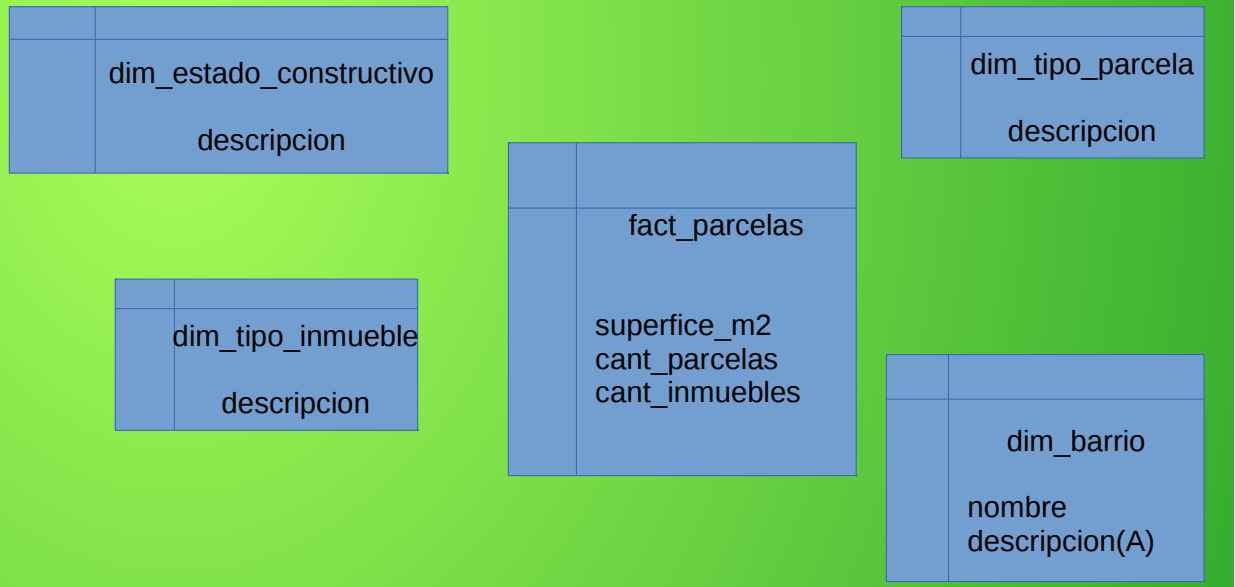
DEMOSTRACION

DIMENSIONES

- Barrio
- Tipo Inmueble
- Estado Inmueble
- Estado Parcela

MEDIDAS

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



PDI

DEMOSTRACION