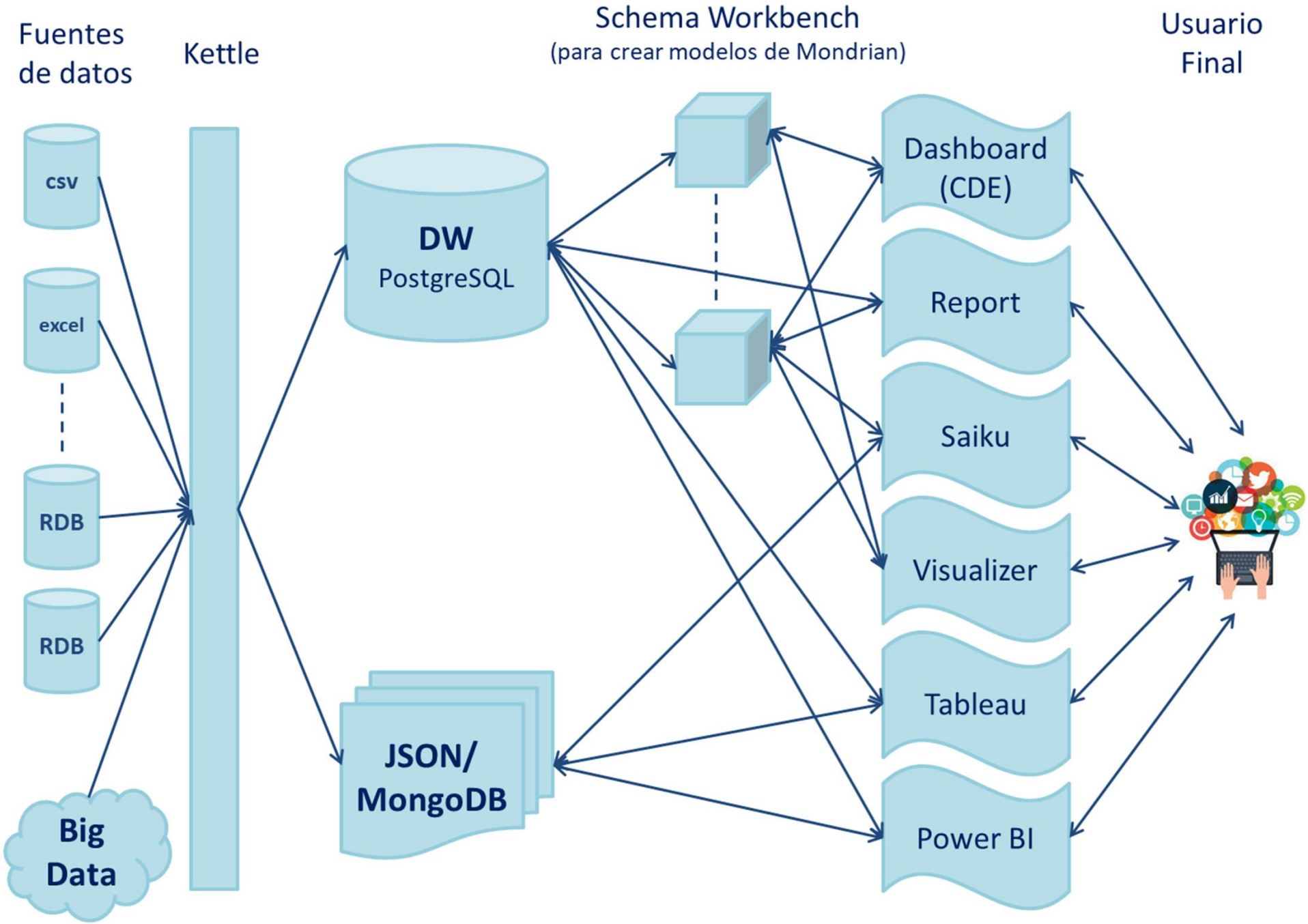
Diseño y Construcción de Data Warehouse

Caso de Estudio

Instituto de Computación - Facultad de Ingeniería

Edición 2021

Integración de las herramientas



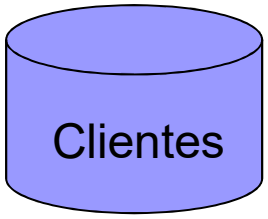
Planteo de la realidad

- Una empresa tiene sucursales en distintas ciudades y abarca la mayoría de los departamentos.

- Requerimientos:
  - Los gerentes desean analizar las ventas.
  - Interesa clasificarlas según los clientes, el tiempo (fecha, mes y año) y familia (tipo) del producto.
  - Interesa evaluar las ventas por ciudad y departamento de los clientes.

Fuentes

- La empresa tiene diferentes formas de almacenamiento de datos.
  - Clientes: Base de Datos
  - Productos: Planilla Excel (.xls)
  - Ventas: Planilla (.csv)



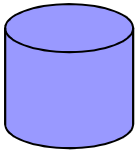
Base de Datos

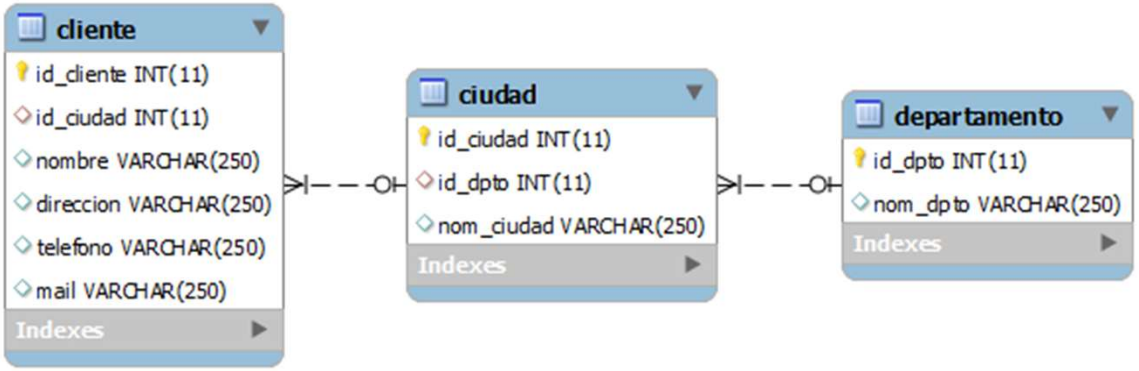
Productos (.xls)

Ventas (.csv)

Fuentes

Clientes

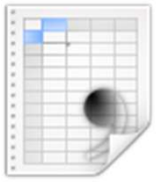




Productos

	A	B	C	D
1	Código del Producto	Nombre del Producto	Código de la Familia	Nombre de la Familia
2	1	Botas	G	Gala
3	2	Buzo	S	Sport
4	3	Camisa	G	Gala
5	4	Chaleco	G	Gala
	...	...	...	...

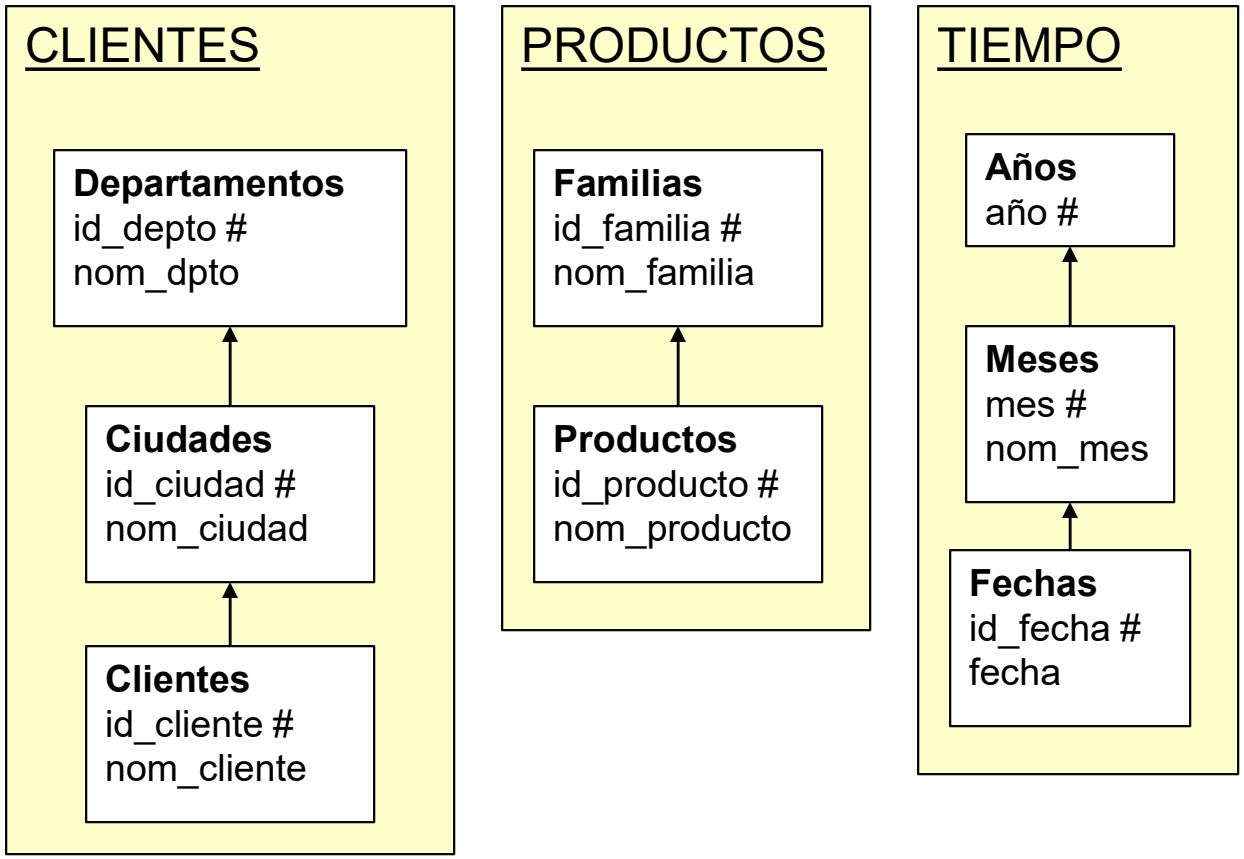
Ventas



	A	B	C	D	E
1	Fecha	Nro. Factura	Cliente	Producto	Cantidad
2	08/10/2007	2636	32	6	1
3	08/10/2007	1166	18	14	1
4	08/10/2007	5061	2	3	2
	...	...	...	...	...

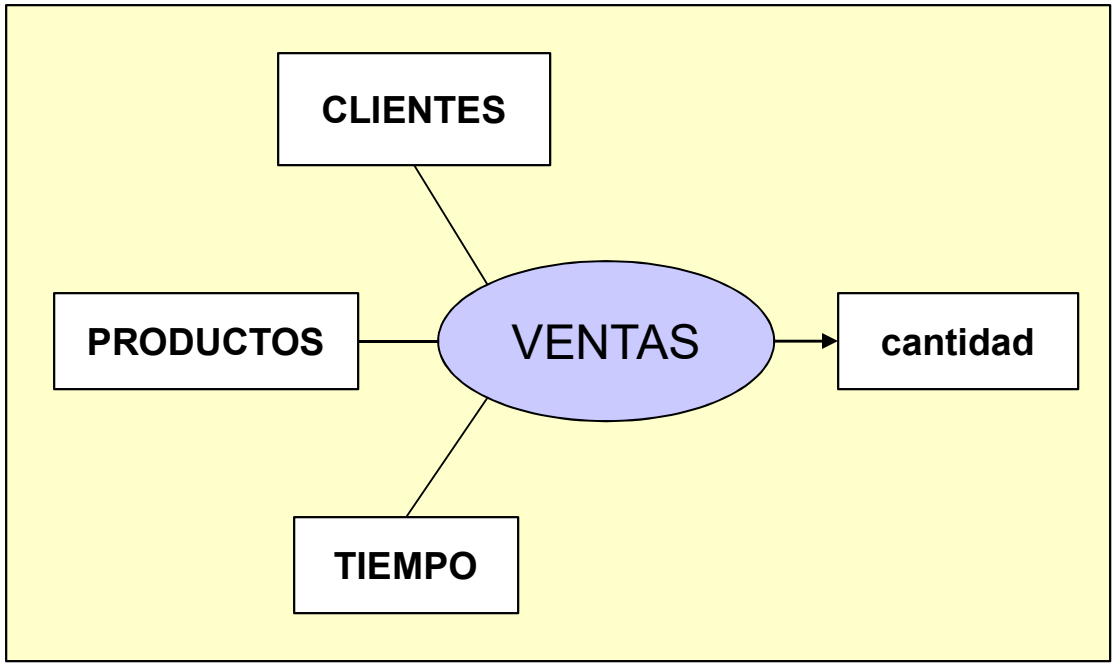
Diseño Conceptual

Dimensiones y Jerarquías



Diseño Conceptual

Relación Dimensional



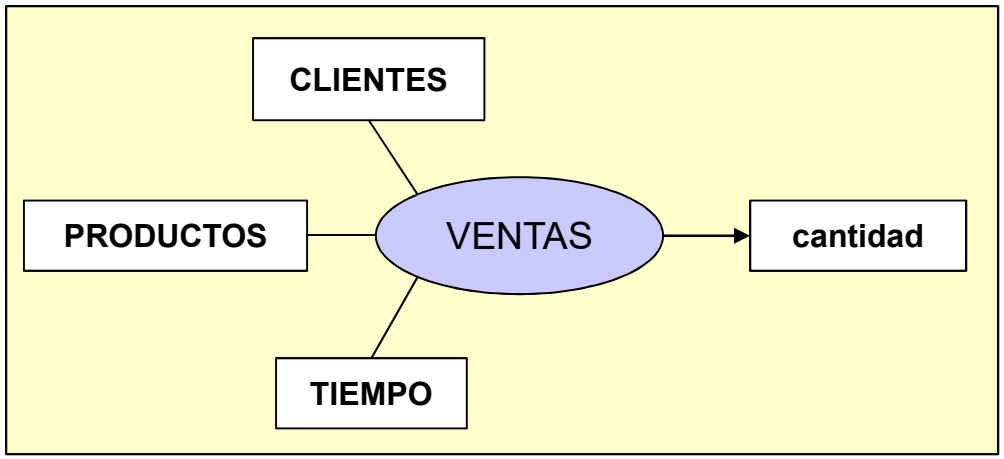
Diseño Conceptual

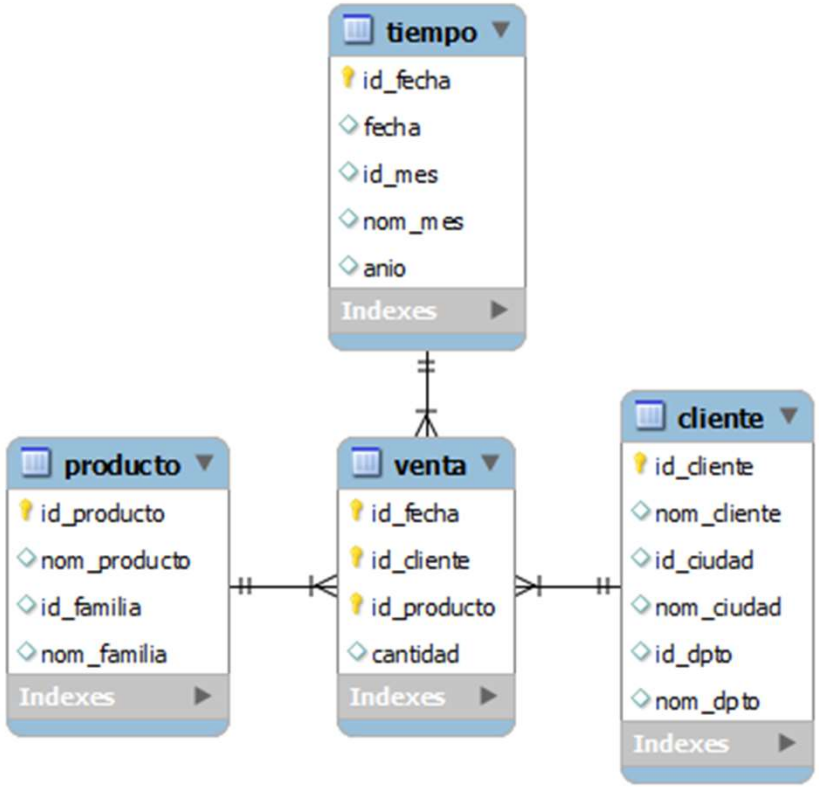
Estudio de Aditividad

Medida / **Dimensión**		Cantidad
Tiempo	**Fechas → Meses**	+
	Meses → Años	+
	Años → ALL	+
Productos	**Productos → Familias**	+
	Familias → ALL	+
Clientes	**Clientes → Ciudades**	+
	Ciudades → Departamentos	+
	Departamentos → ALL	+

Diseño Lógico

Esquema Estrella





Diseño Físico

- Seleccionar los tipos de datos más pequeños del DBMS que permitan representar los datos.

Name	Storage Size	Description	Range
smallint	2 bytes	small-range integer	-32768 to +32767
integer	4 bytes	typical choice for integer	-2147483648 to +2147483647
bigint	8 bytes	large-range integer	-9223372036854775808 to +9223372036854775807
decimal	variable	user-specified precision, exact	up to 131072 digits before the decimal point; up to 16383 digits after the decimal point
numeric	variable	user-specified precision, exact	up to 131072 digits before the decimal point; up to 16383 digits after the decimal point
real	4 bytes	variable-precision, inexact	6 decimal digits precision
double precision	8 bytes	variable-precision, inexact	15 decimal digits precision
smallserial	2 bytes	small autoincrementing integer	1 to 32767
serial	4 bytes	autoincrementing integer	1 to 2147483647
bigserial	8 bytes	large autoincrementing integer	1 to 9223372036854775807

postgreSQL 10 – Tipos numéricos

https://www.postgresql.org/docs/current/static/datatype-numeric.html

Diseño Físico

- Utilizar claves subrogadas (artificiales) de tipo entero sin signo para identificar los hechos y los diferentes niveles de las dimensiones.

- Caso particular dimensión **Tiempo**. Claves subrogadas con formato `YYYYMMDD` permiten:
  - a partir de una fecha generar el identificador y viceversa
  - preservan la relación de orden

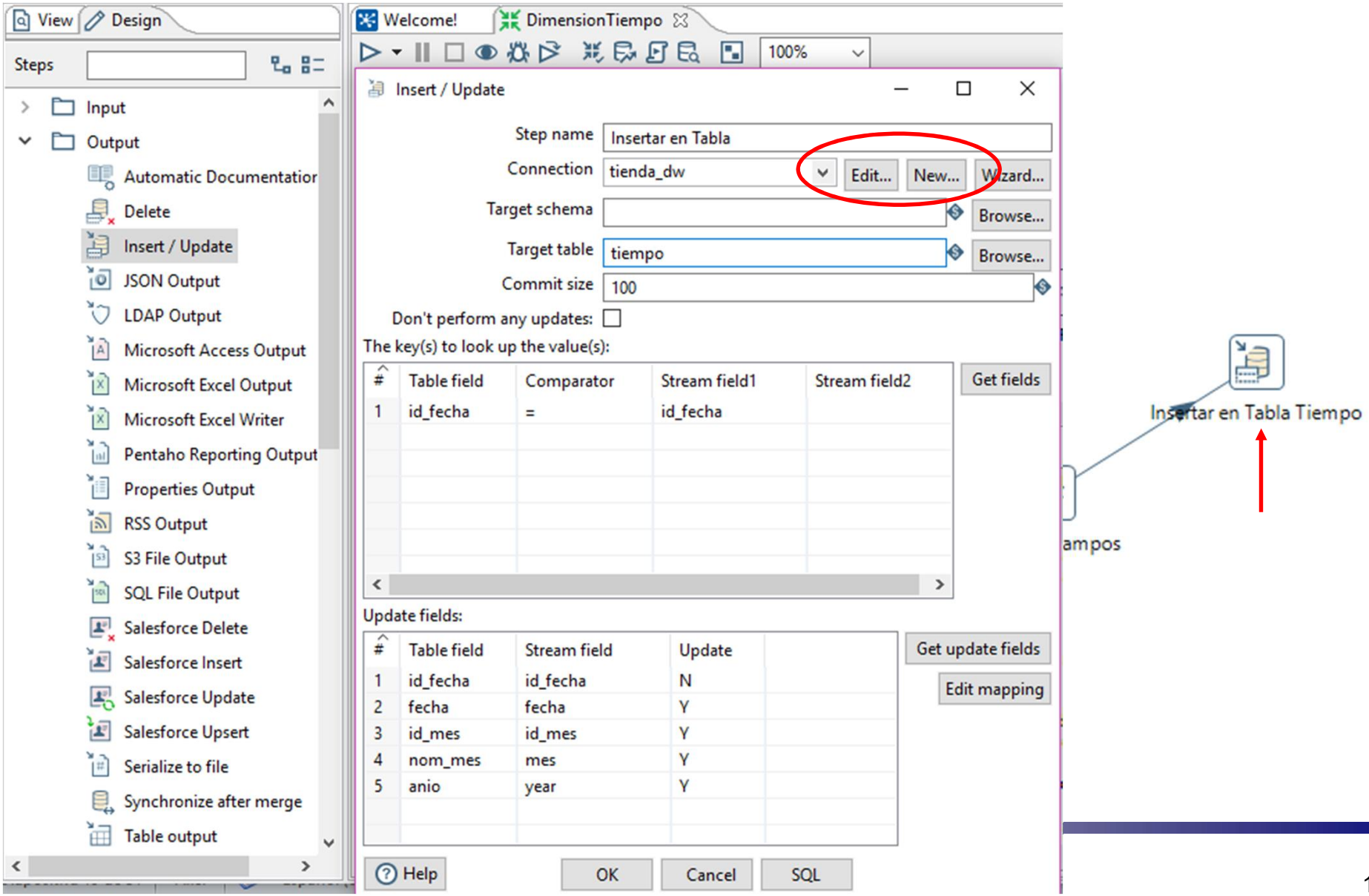
Diseño Físico

- Seleccionar un conjunto de índices adecuado, que permita realizar joins de forma eficiente: B-TREE, HASH, etc.
- Estudiar características particulares del DBMS para aprovecharlo al máximo. Por ejemplo en PostgreSQL:
  - Table partitioning: Divide tablas muy grandes en pequeñas tablas (https://www.postgresql.org/docs/current/ddl-partitioning.html)
  - VACUUM: Recupera el almacenamiento ocupado por «tuplas muertas» (https://www.postgresql.org/docs/current/static/sql-vacuum.html)
  - ANALIZE: Recopila estadísticos de la BD (https://www.postgresql.org/docs/current/static/sql-analyze.html)
  - REINDEX: Reconstruye índices (https://www.postgresql.org/docs/current/static/sql-reindex.html)

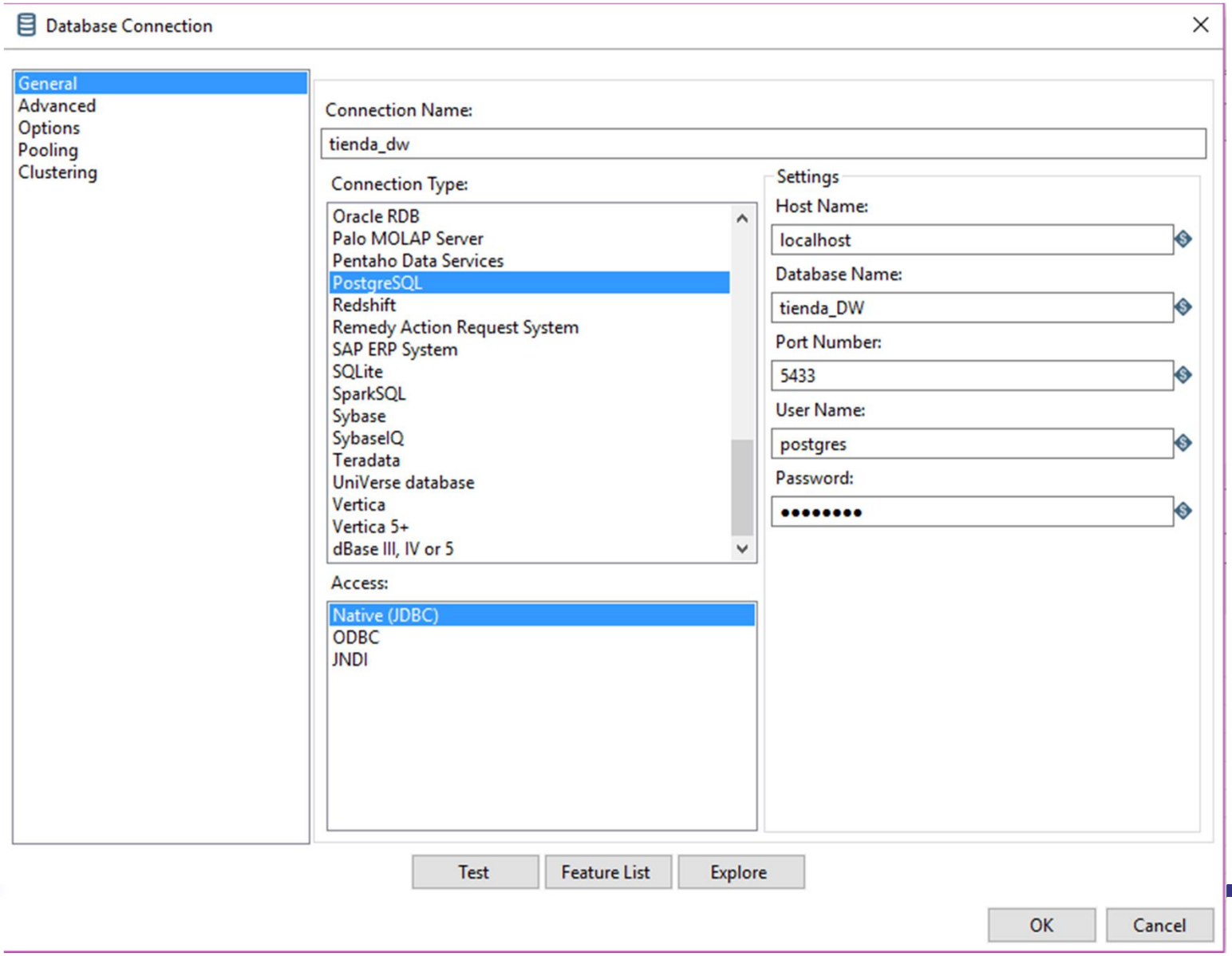
Carga de Dimensiones

- Dimensión *Tiempo*:
  - Generar las fechas
  - Para cada fecha generar los identificadores y los restantes campos de la tabla.
  - Insertar en la tabla *Tiempo* del DW.

Carga de Dimensión: Tiempo

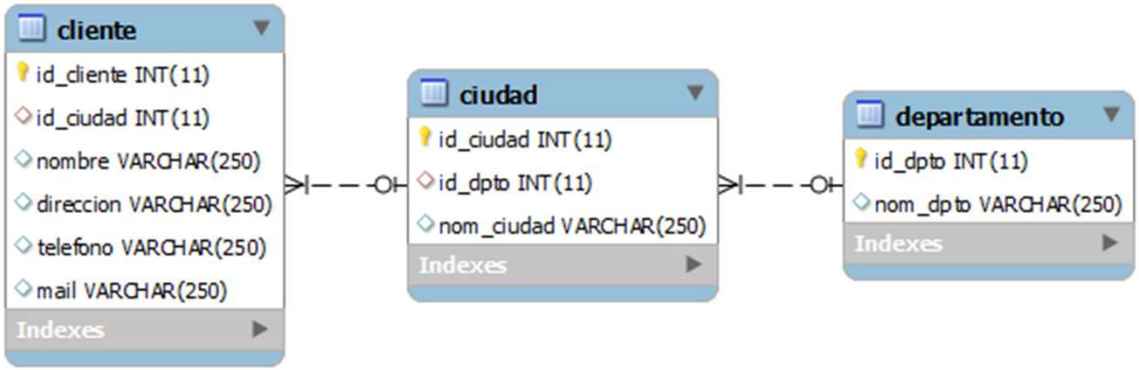


Carga de Dimensión: Tiempo

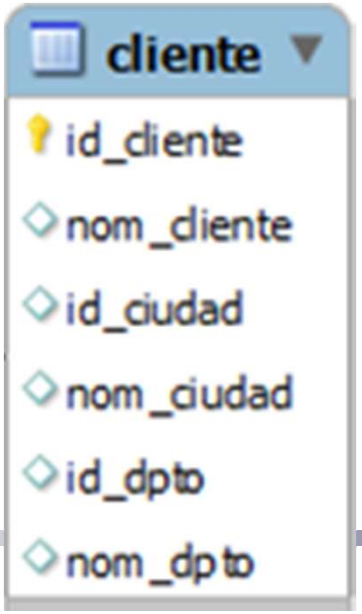


Carga de Dimensión: Cliente

- BD Fuente (BD):



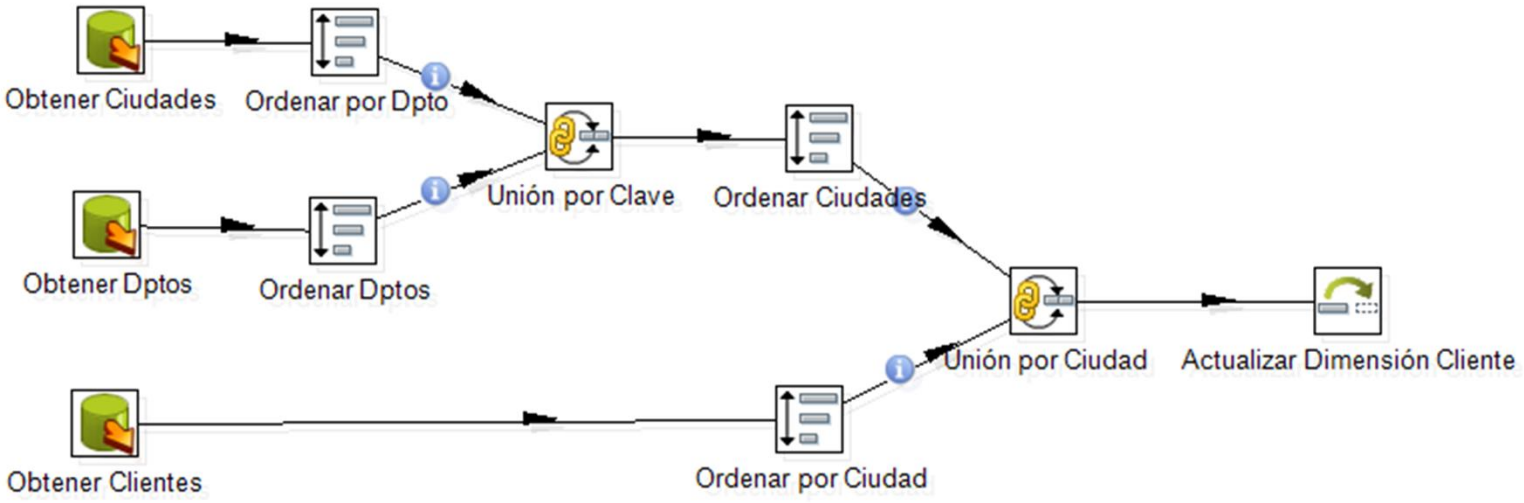
- Tabla en el DW:



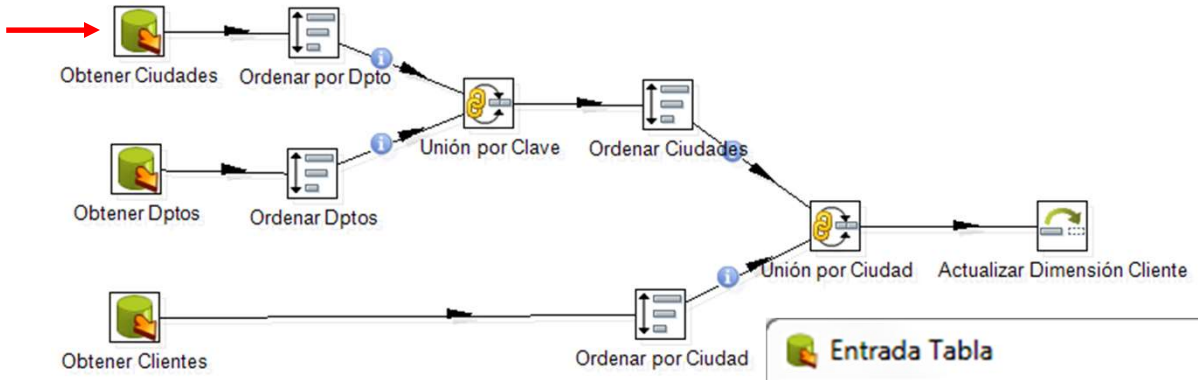
Carga de Dimensión: Cliente

- Dimensión *Cliente*:
  - Leer los registros de las tres tablas de la base fuente.
  - Unir ("join") los registros de cliente, ciudad y dpto.
  - Insertar en la tabla CLIENTE del DW.

- Cliente: Transformación

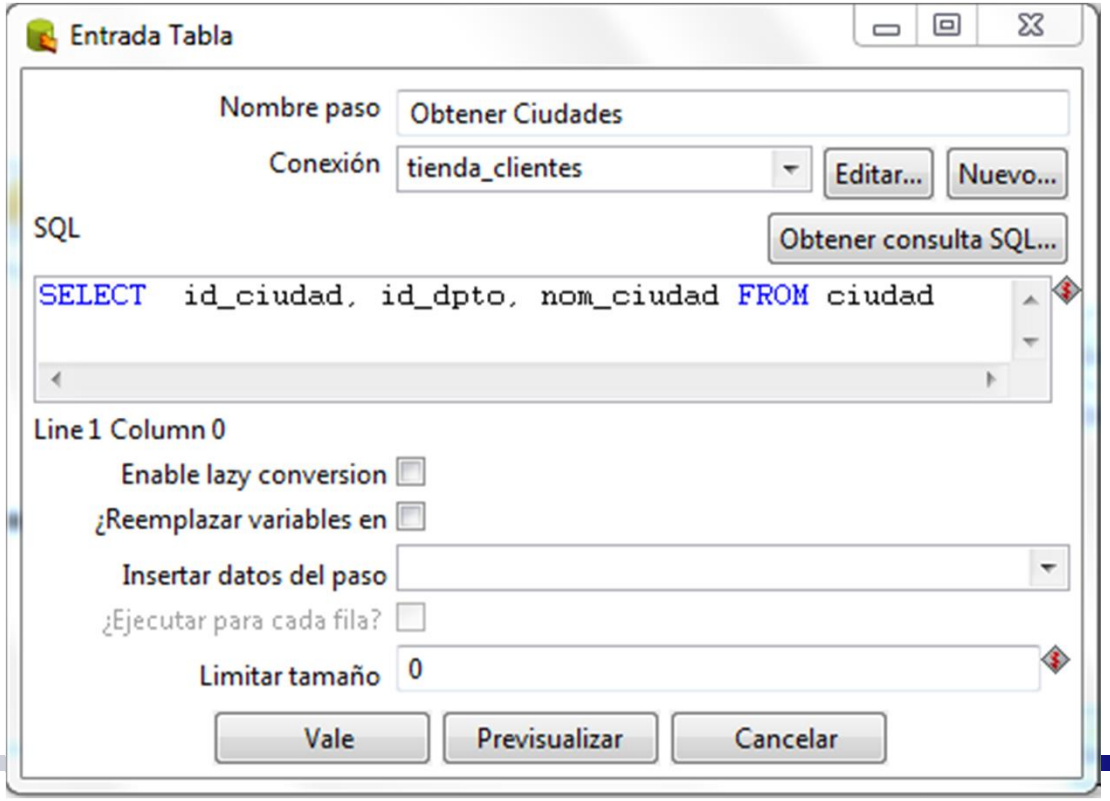


Carga de Dimensión: Cliente

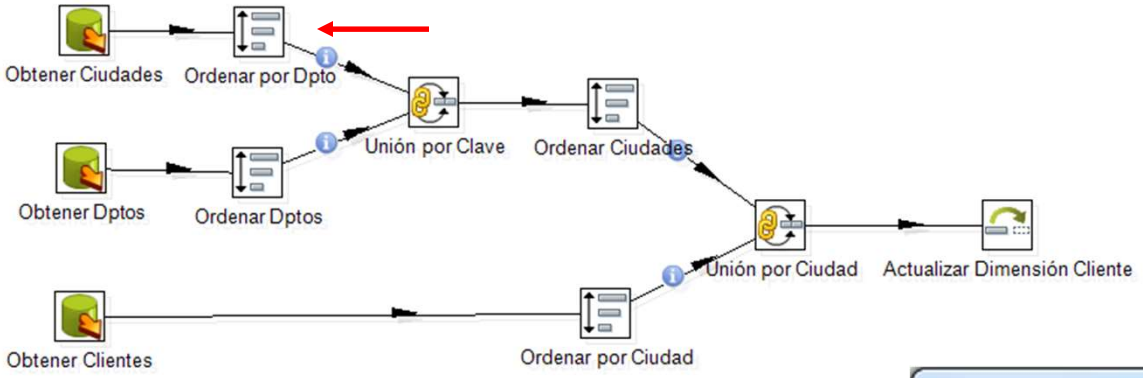


Cliente: *Step*

«Obtener ciudades»

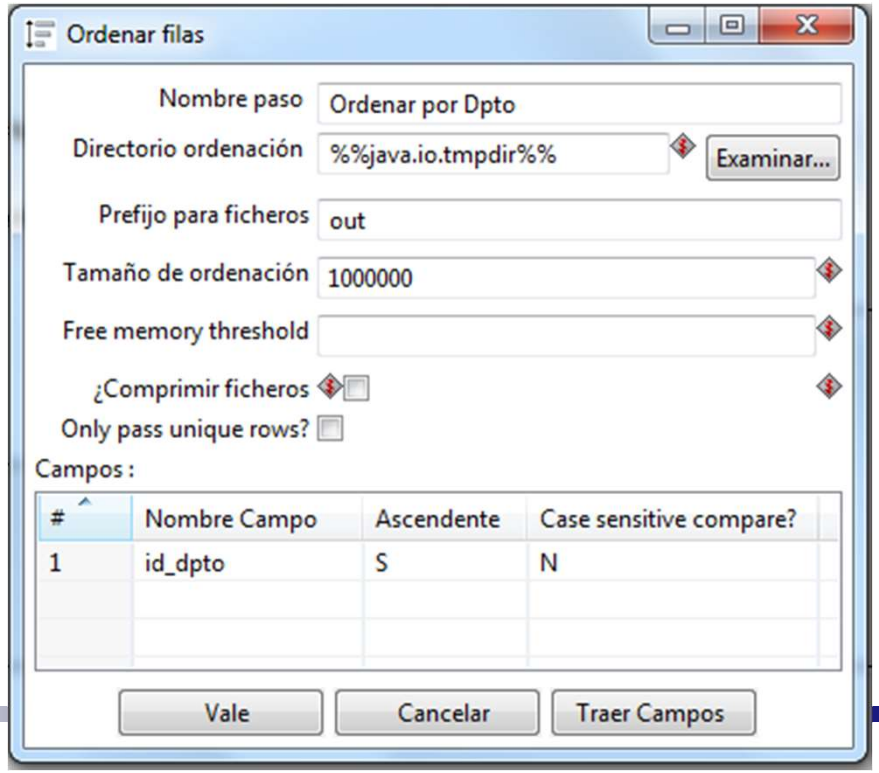


Carga de Dimensión: Cliente

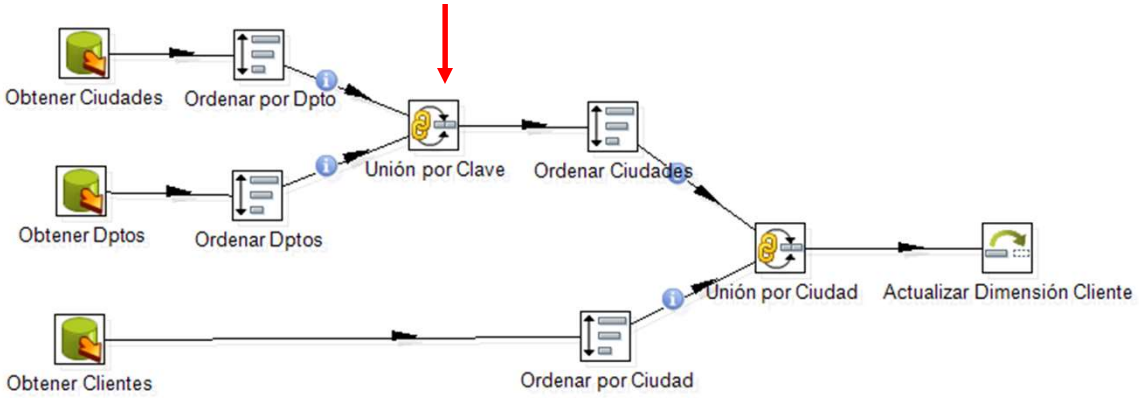


Cliente: *Step*

«Ordenar por Dpto.»

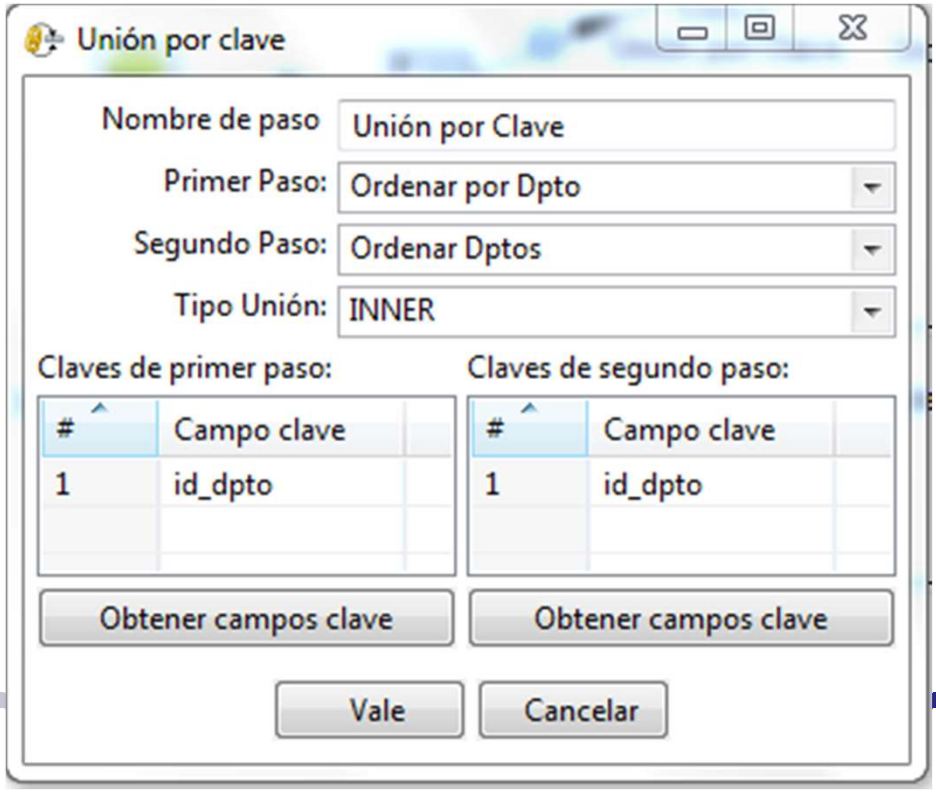


Carga de Dimensión: Cliente

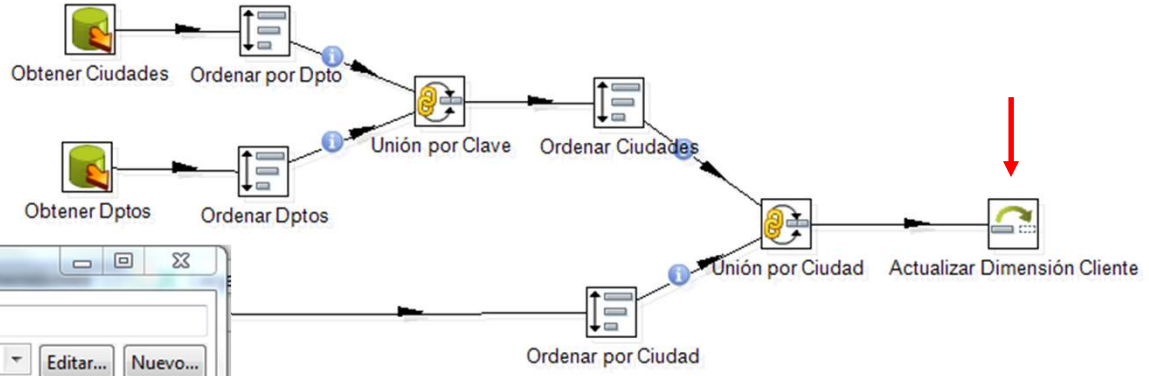


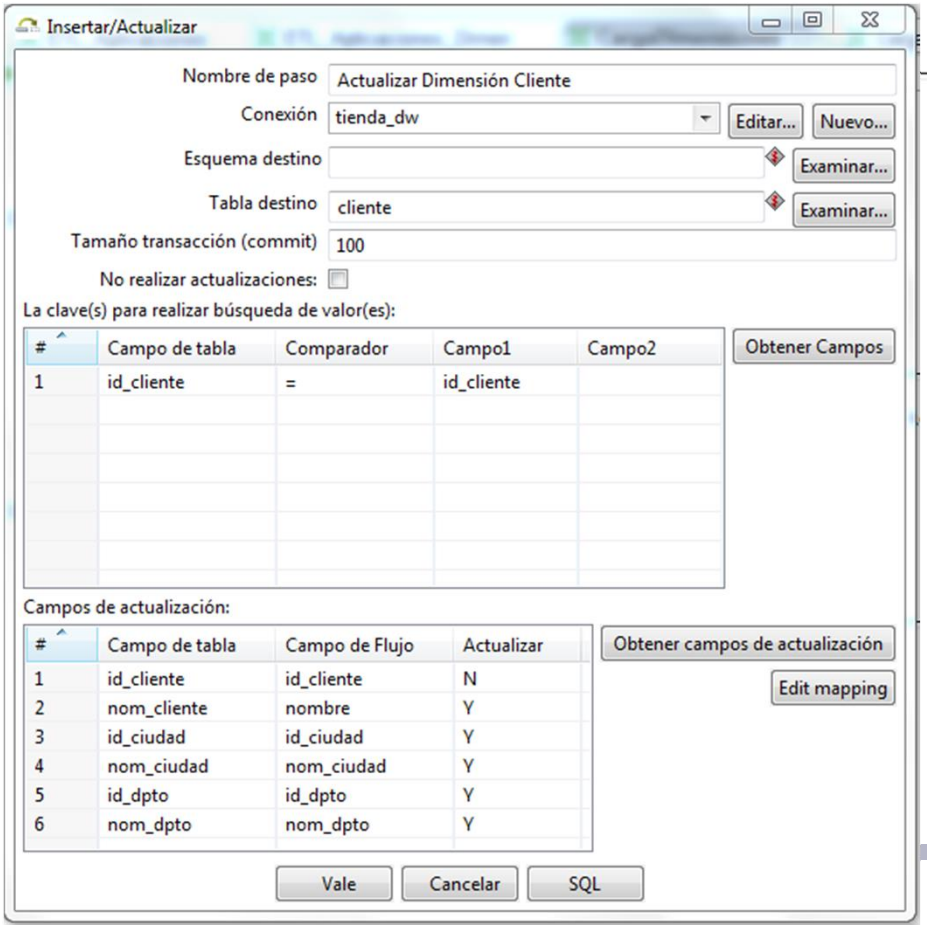
Cliente: *Step*

«Unión por Clave»



Carga de Dimensión: Cliente





Cliente: *Step*

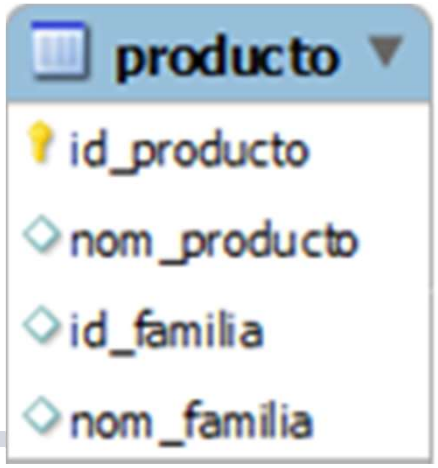
«Actualizar dimensión Cliente»

Carga de Dimensión: Producto

- Fuente (.xls):

	A	B	C	D
1	Código del Producto	Nombre del Producto	Código de la Familia	Nombre de la Familia
2	1	Botas	G	Gala
3	2	Buzo	S	Sport
4	3	Camisa	G	Gala
5	4	Chaleco	G	Gala
	...	...	...	...

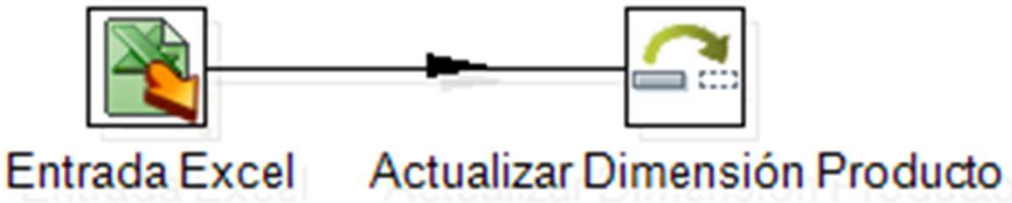
- Tabla en el DW:



Carga de Dimensión: Producto

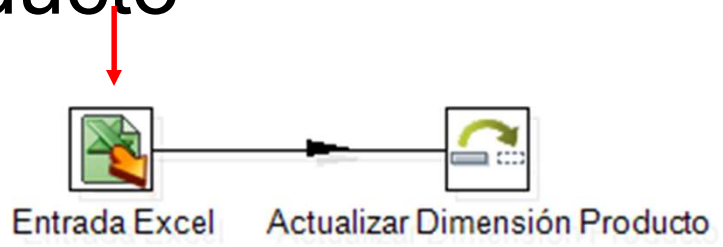
- Dimensión *Producto*
  - Leer los registros de la planilla Excel (.xls)
  - Insertar en la tabla Producto del DW

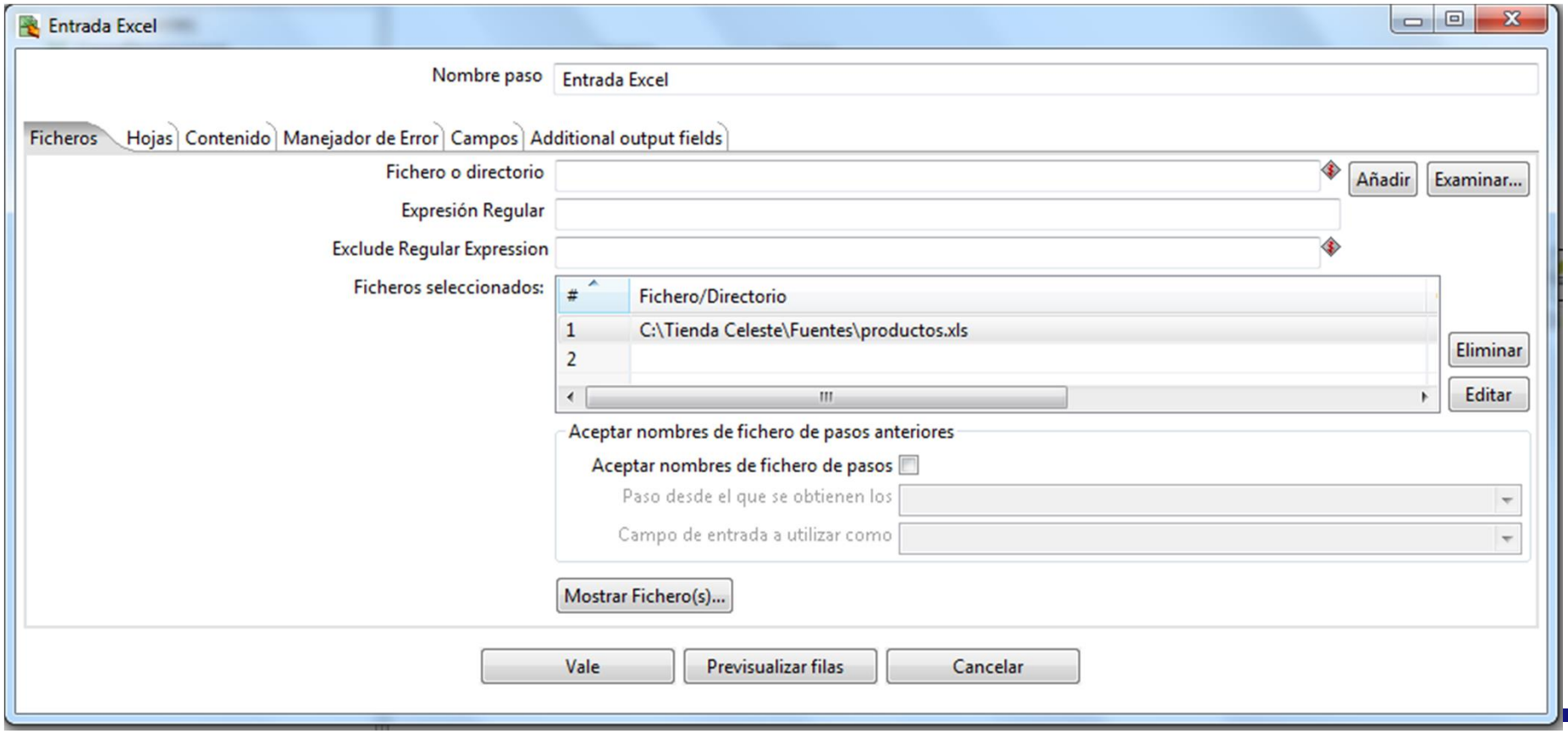
- Transformación



Carga de Dimensión: Producto

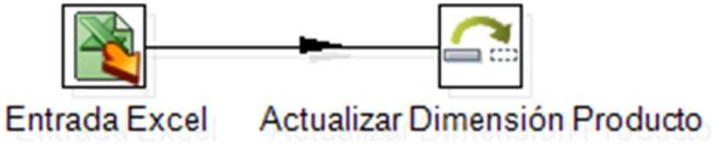
Producto: *Step «Entrada Excel»*

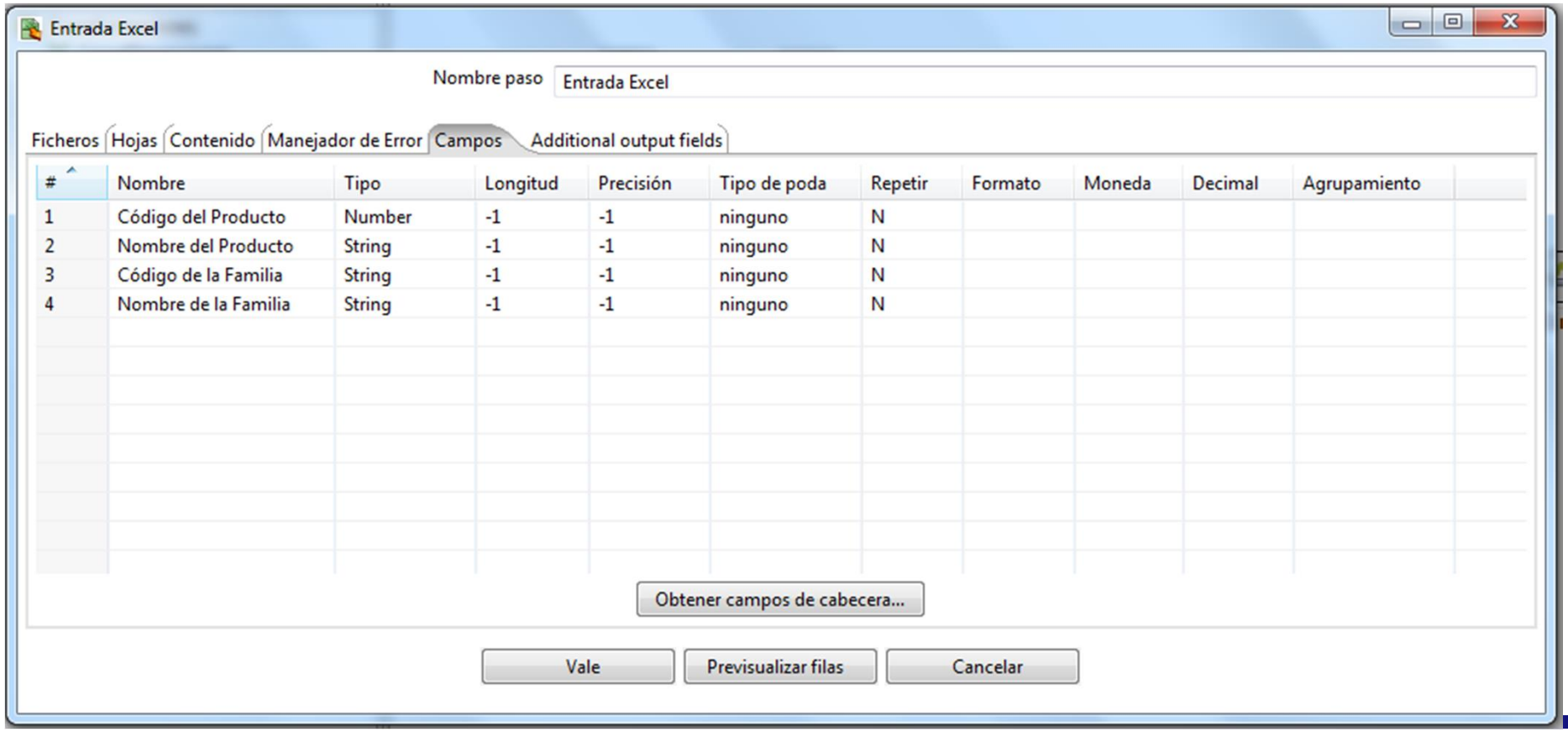




Carga de Dimensión: Producto

Producto: *Step «Entrada Excel»*

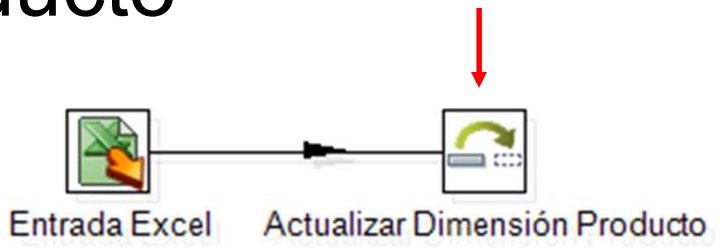


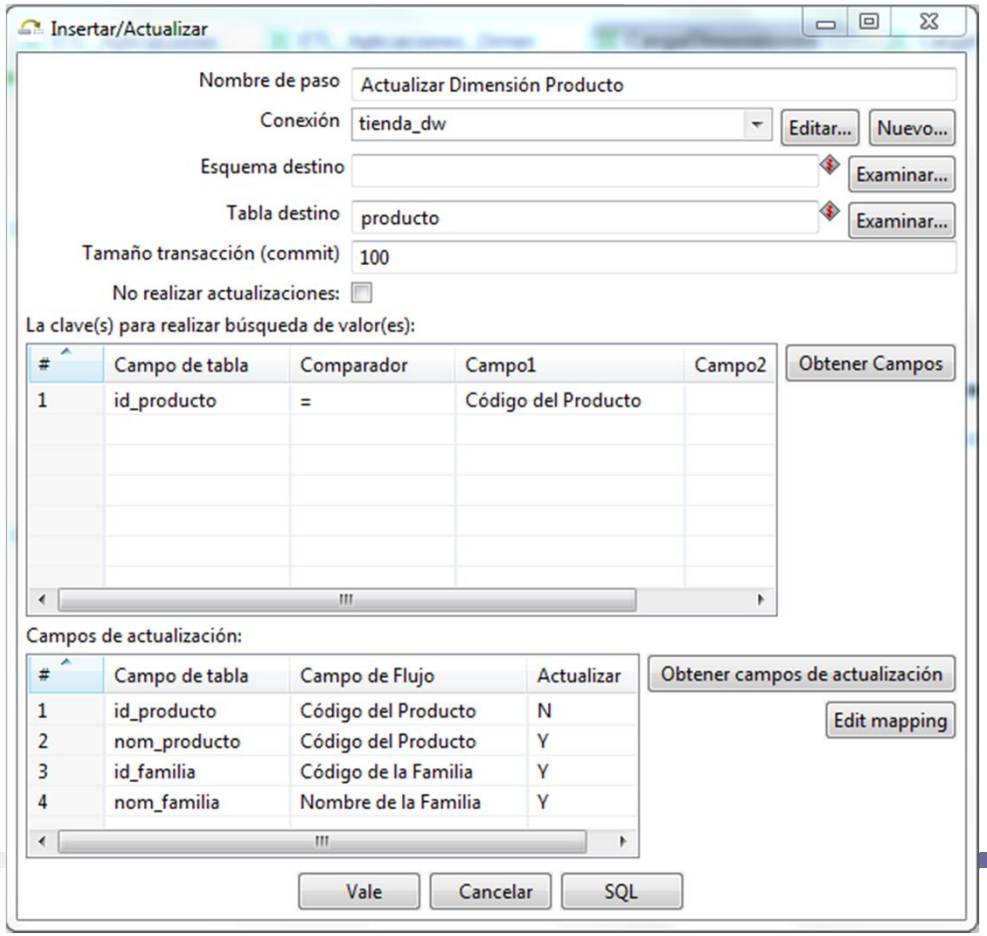


#	Nombre	Tipo	Longitud	Precisión	Tipo de poda	Repetir	Formato	Moneda	Decimal	Agrupamiento
1	Código del Producto	Number	-1	-1	ninguno	N				
2	Nombre del Producto	String	-1	-1	ninguno	N				
3	Código de la Familia	String	-1	-1	ninguno	N				
4	Nombre de la Familia	String	-1	-1	ninguno	N				

Carga de Dimensión: Producto

Producto: *Step «Actualizar Dimensión Producto»*





Carga de Hechos

- Ventas
  - Fuente (.csv):

	A	B	C	D	E
1	Fecha	Nro. Factura	Cliente	Producto	Cantidad
2	08/10/2007	2636	32	6	1
3	08/10/2007	1166	18	14	1
4	08/10/2007	5061	2	3	2
	...	...	...	...	...

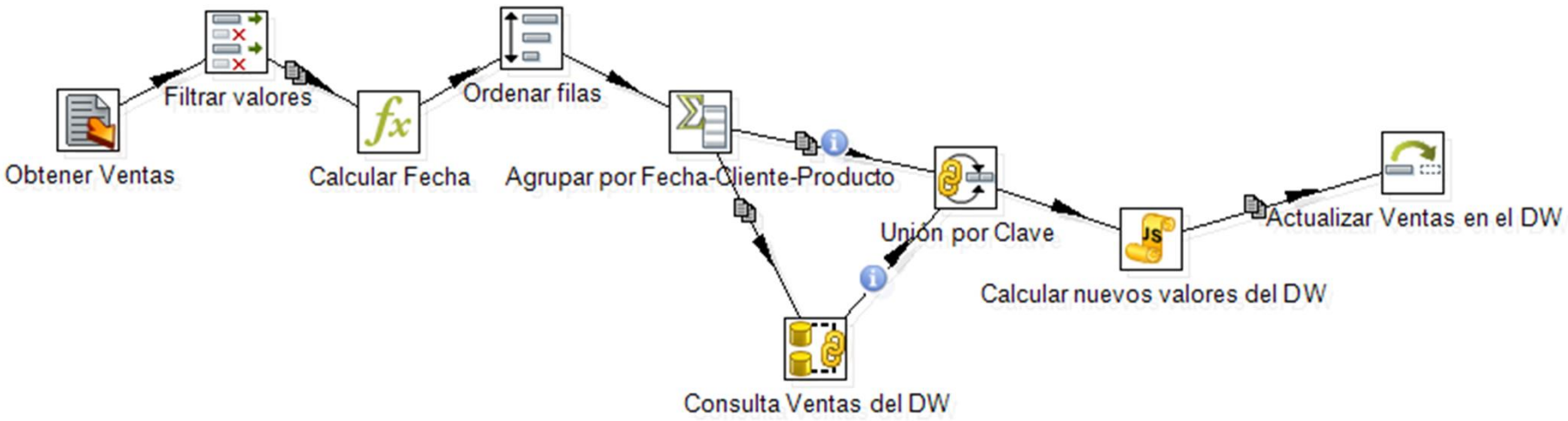
  - Tabla de Hechos en el DW:

Carga de Hechos

- Tabla de Hechos *Ventas*
  - Leer los registros de la planilla de Ventas.
  - Filtrar los campos que realmente son necesarios.
  - Calcular *id_fecha* (clave foránea a dimensión Tiempo).
  - Agrupar por Fecha-Cliente-Producto.
  - Para las combinaciones Fecha-Cliente-Producto que ya existen en el DW, sumar la cantidad.
  - Insertar en la tabla de hechos *Ventas* del DW.

Carga de Hechos: Ventas

- Transformación



Leer datos

Filtrar

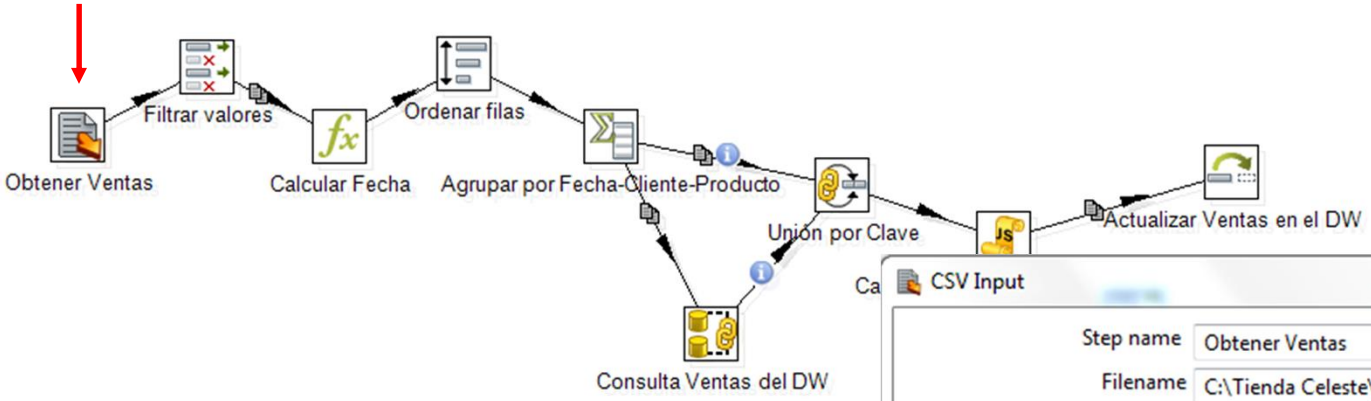
Calcular ID Fecha

Agrupar por Fecha-Cliente-Producto

Verificar si la combinación ya existe en el DW y Sumar

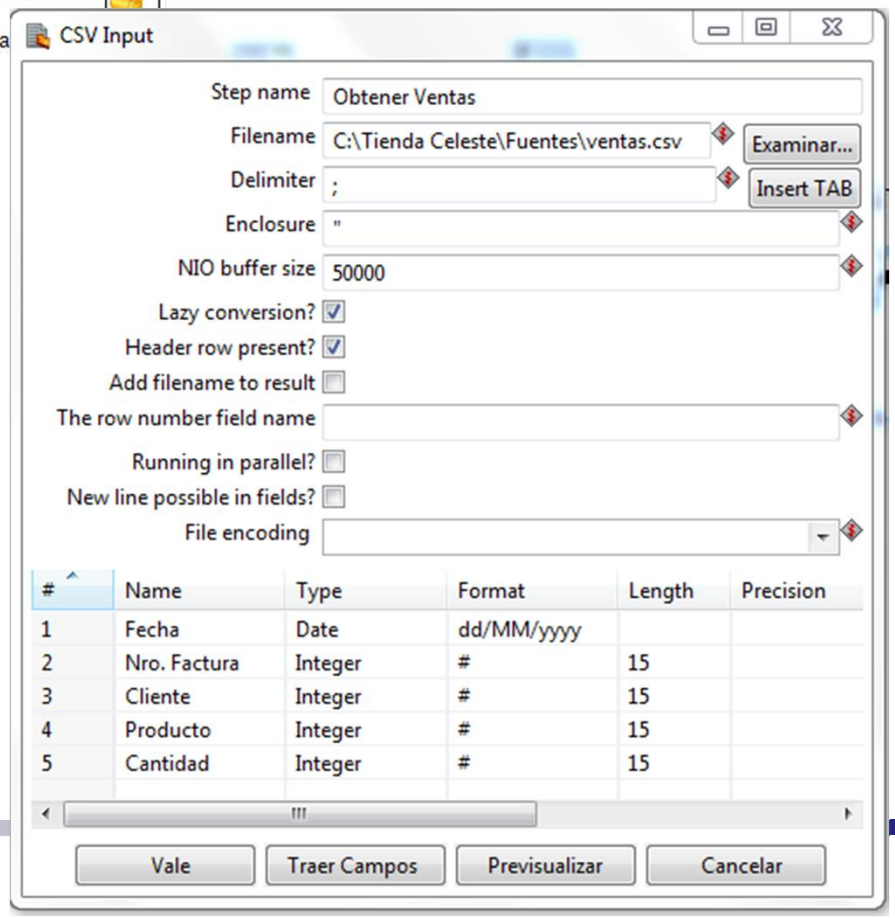
Actualizar tabla de Hechos

Carga de Hechos



Ventas: *Step*

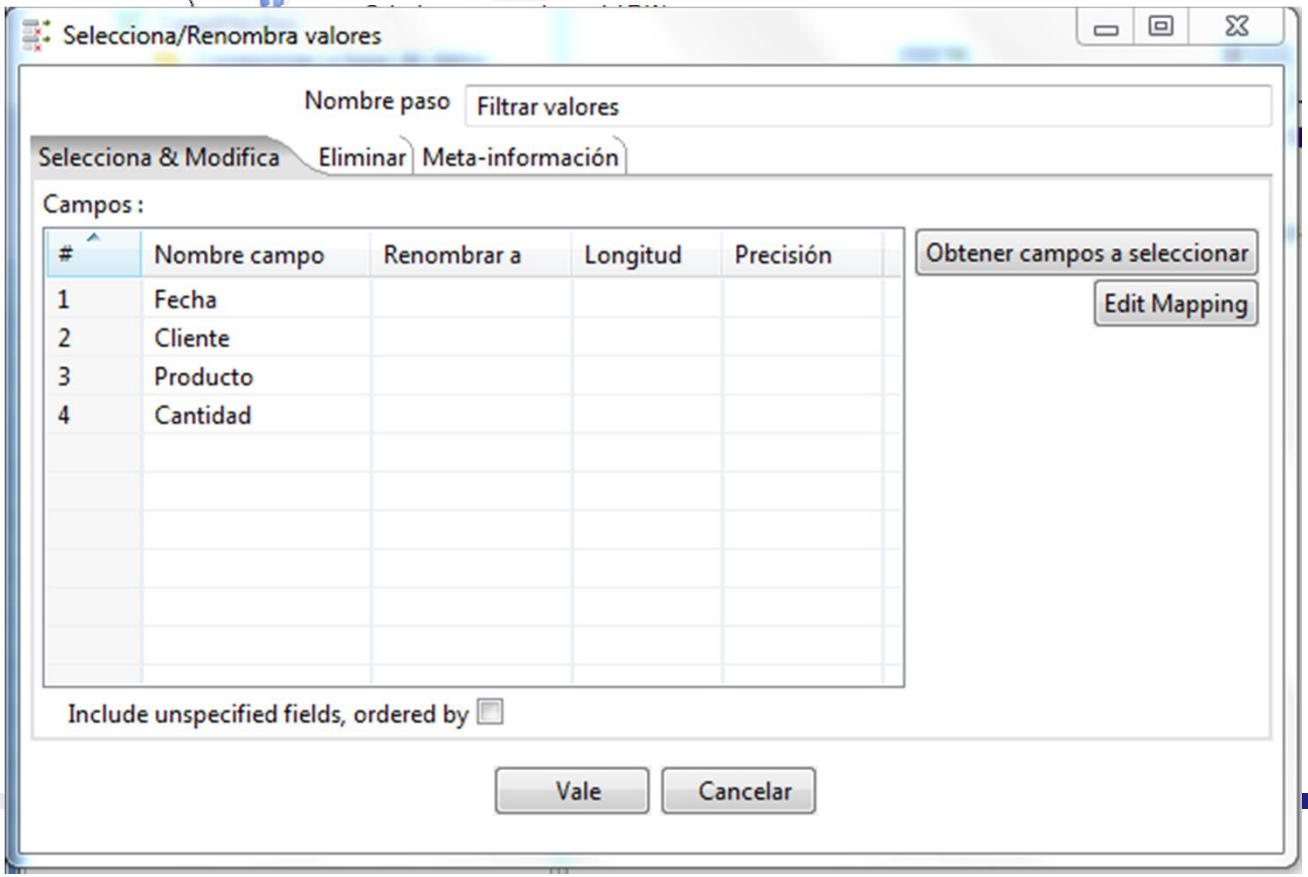
«Obtener Ventas»



Carga de Hechos

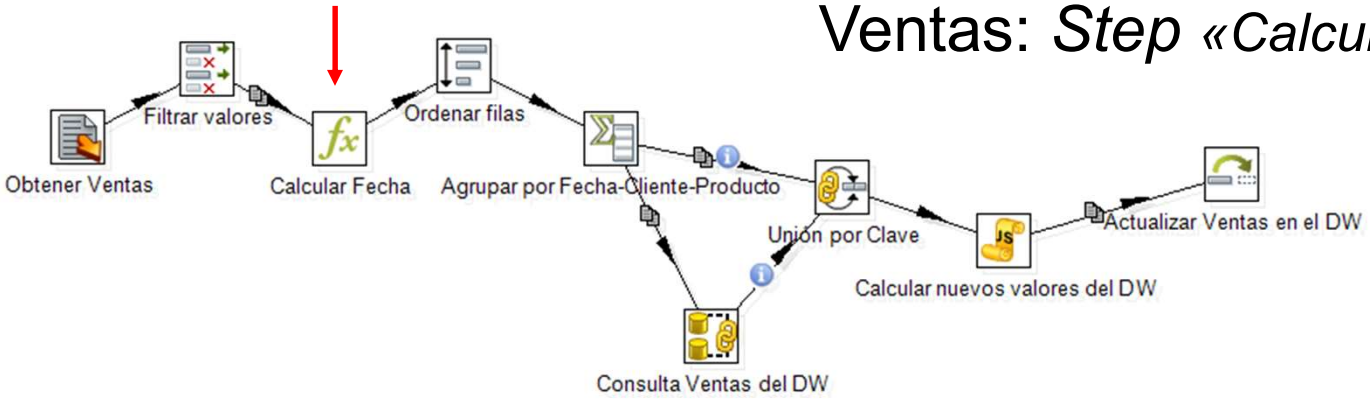
Ventas: *Step*

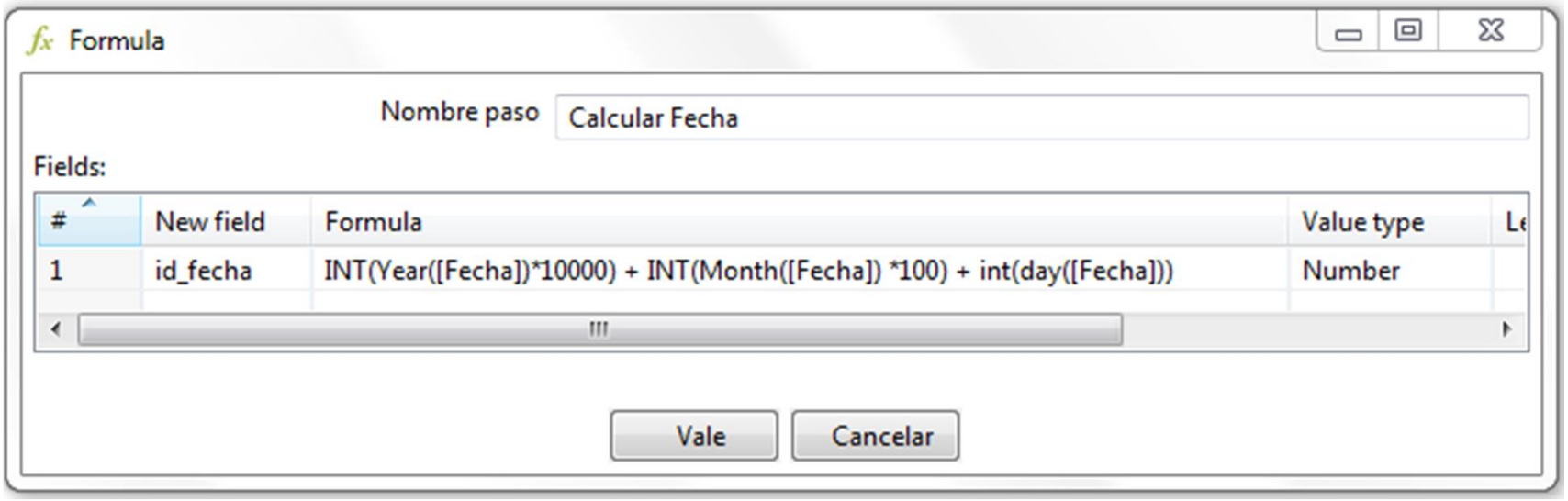
«Filtrar valores»



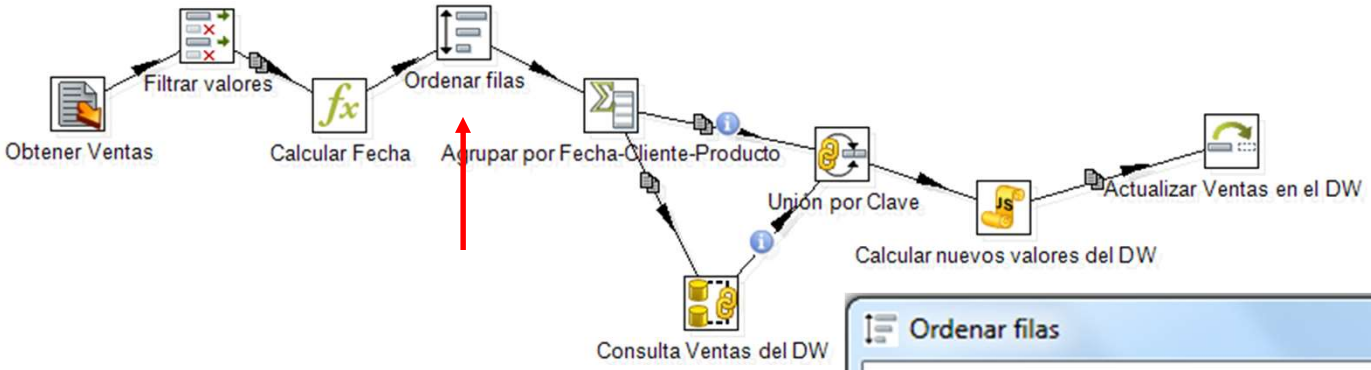
Carga de Hechos

Ventas: *Step «Calcular Fecha»*



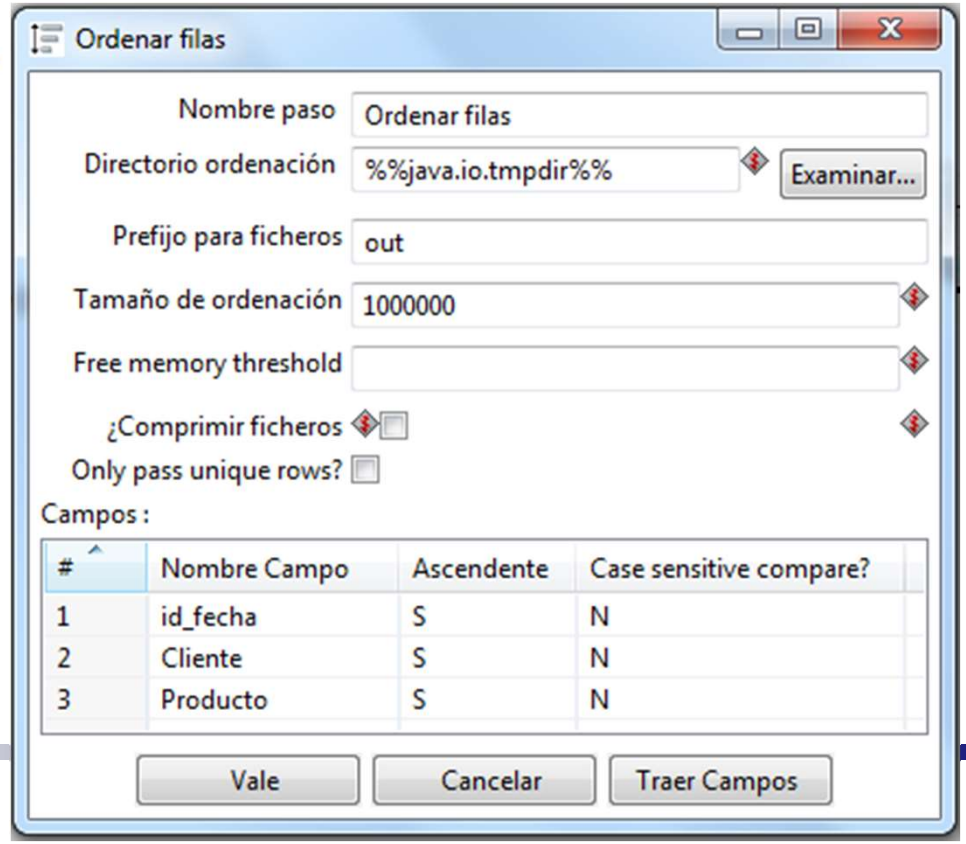


Carga de Hechos

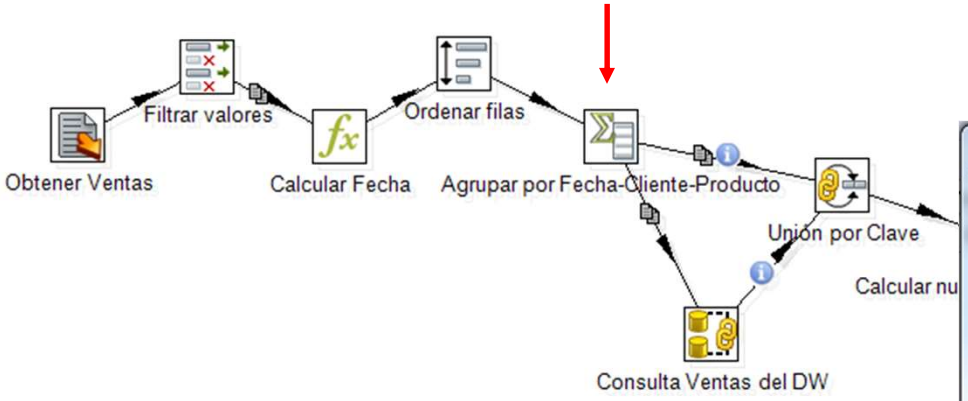


Ventas: *Step*

«Ordenar filas»

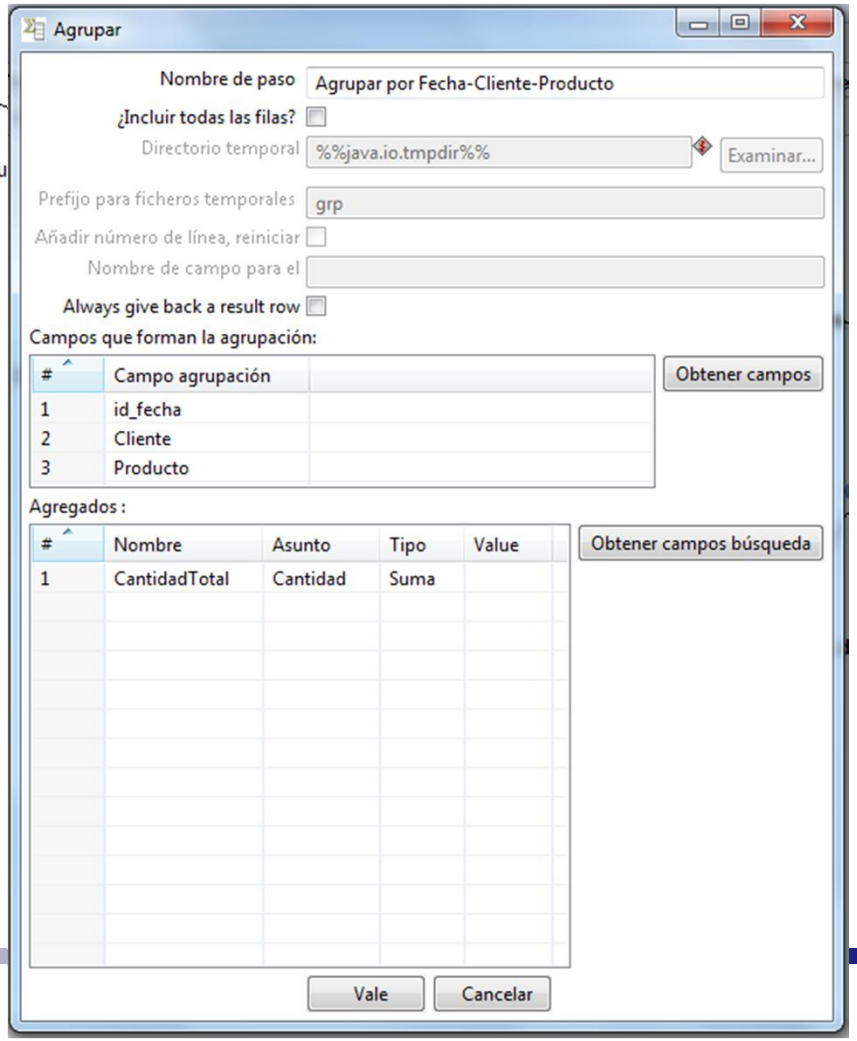


Carga de Hechos

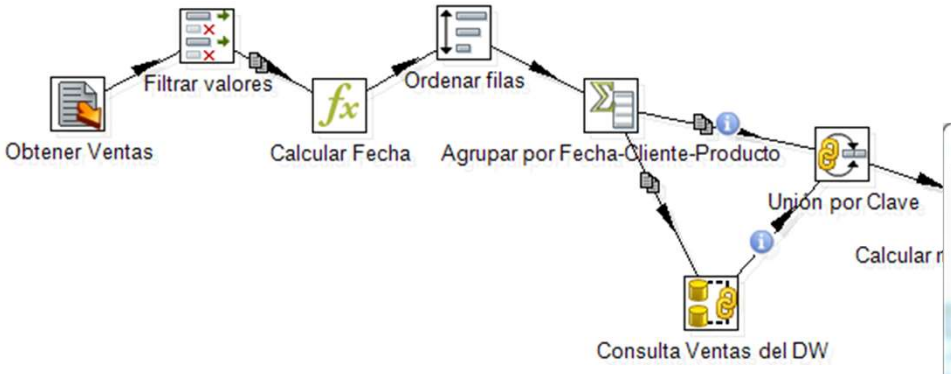


Ventas: *Step*

«Agrupar por Fecha-Cliente-Producto»

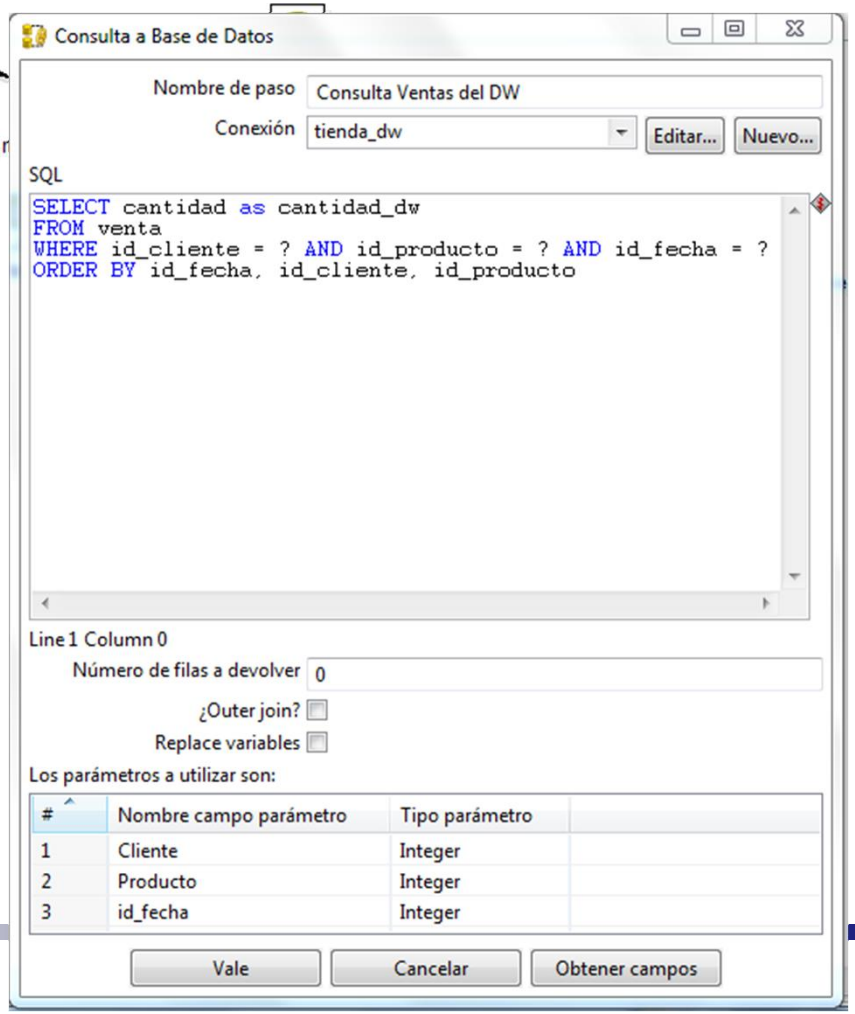


Carga de Hechos



Ventas: *Step*

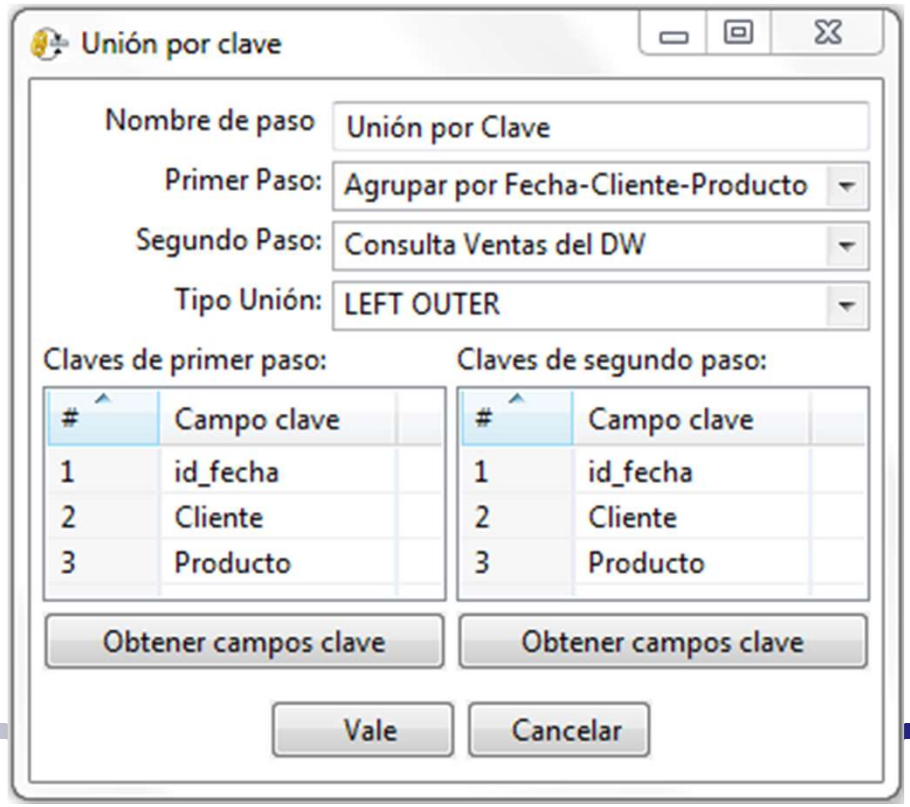
«Consultas Ventas del DW»



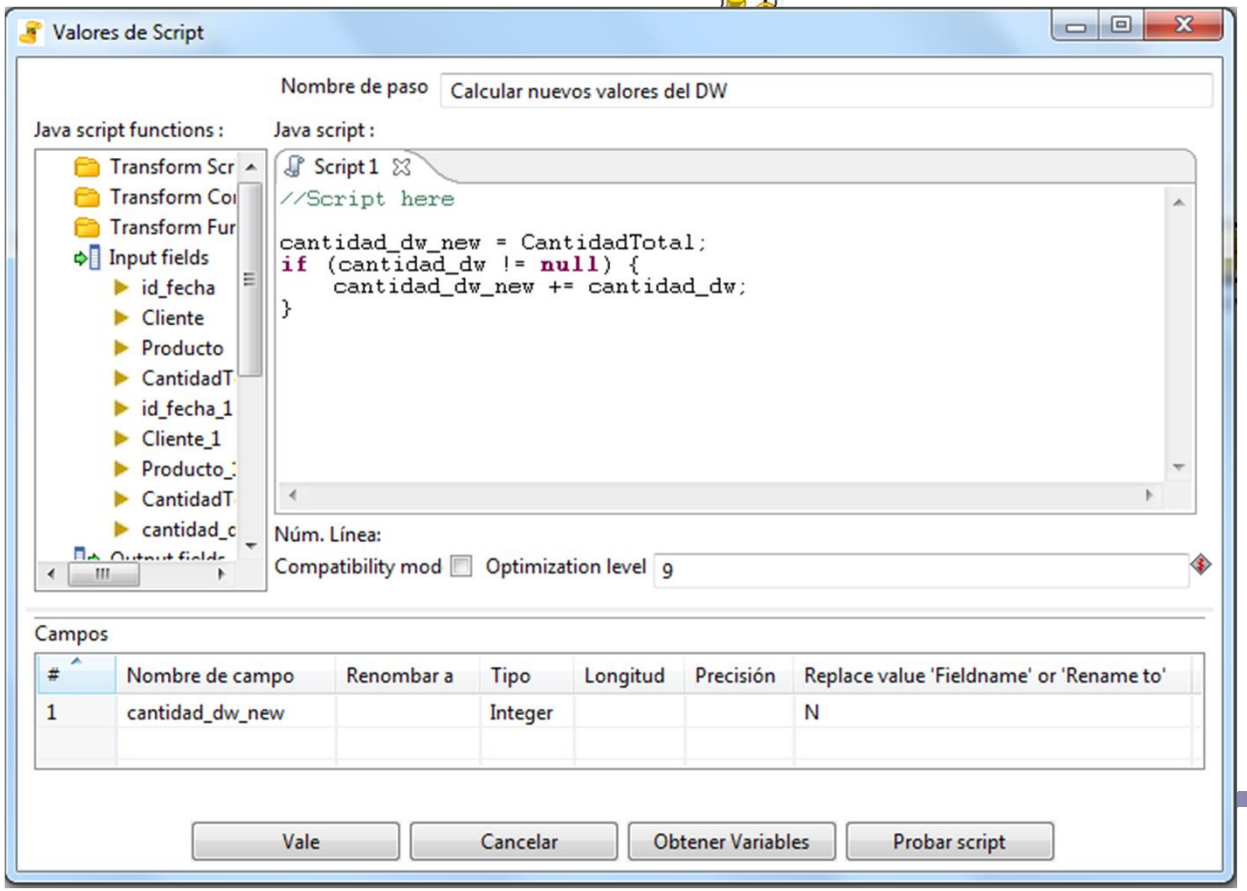
Carga de Hechos

Ventas: *Step*

«Unión por Clave»



Carga de Hechos

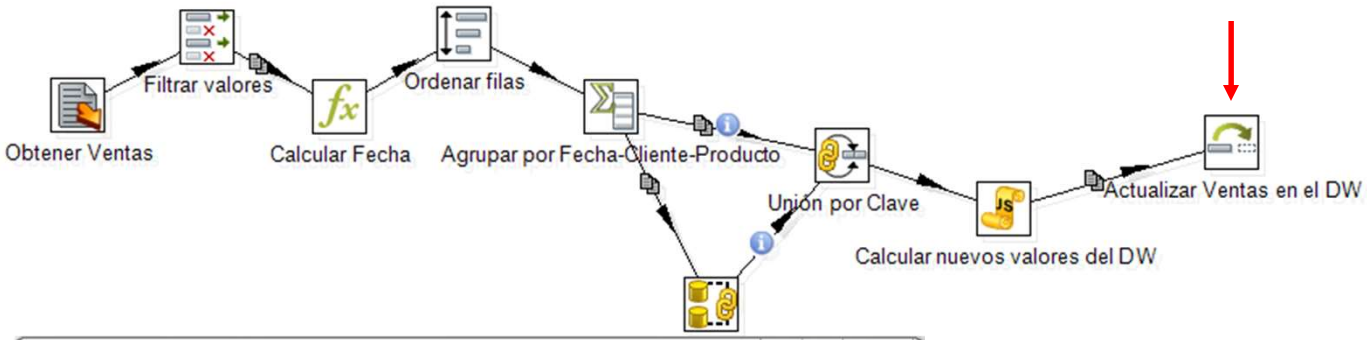


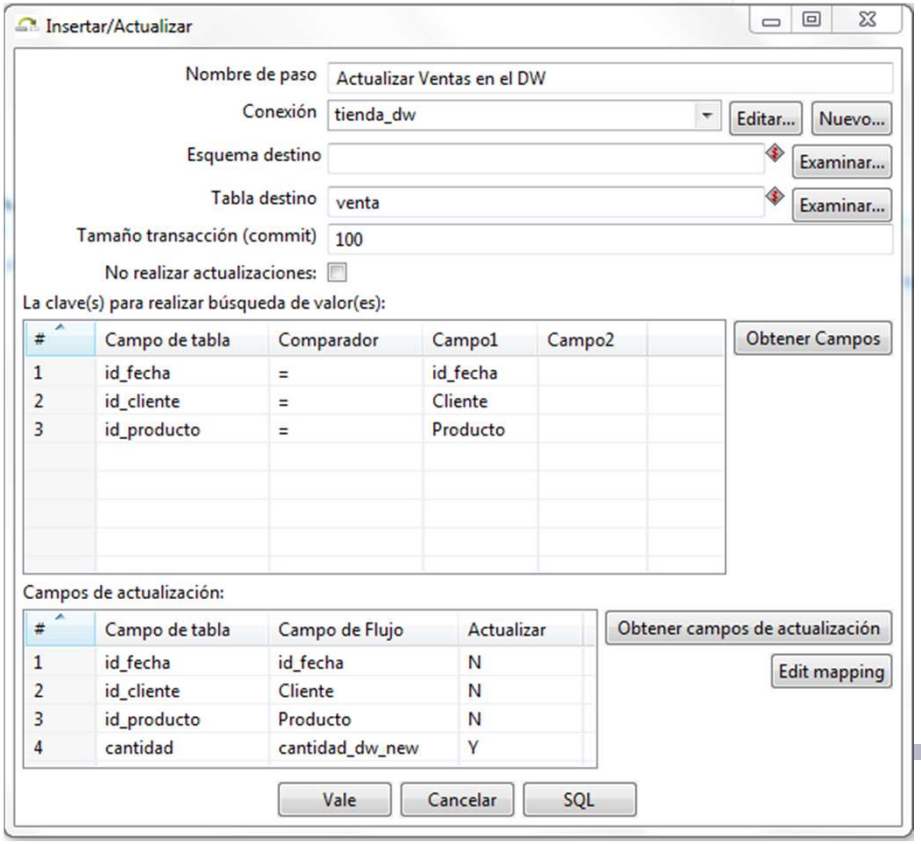
Ventas: *Step*

«Calcular nuevos valores del DW»

Para las combinaciones Fecha-Cliente-Producto que ya existen en el DW, sumar la cantidad

Carga de Hechos





Ventas: *Step*

«Actualizar Ventas en el DW»

Transformaciones

- En general, la dimensión *Tiempo* podría ser cargada solo la primera vez, eventualmente cada mucho tiempo.

- Dimensiones *Productos* y *Clientes*:
  - Varían más frecuentemente, dimensiones dinámicas.
  - Deberían actualizarse en el DW cada vez que se van a cargar los hechos.

Trabajo (Job)

- Primera Carga:
  - Cargar todas las Dimensiones
  - Cargar los Hechos

- Actualización del DW:
  - Cargar las Dimensiones “dinámicas” *Clientes y Productos*
  - Cargar los Hechos *Ventas*