Diseño y Construcción de Data Warehouse

Instituto de Computación - Facultad de Ingeniería

Edición 2017

Pentaho: Ejemplo de Diseño y Carga

- Temario
  - Planteo de la realidad
  - Fuentes
  - Diseño Conceptual
  - Diseño Lógico
  - Diseño Físico
  - Carga de Dimensiones
  - Carga de Hechos
  - Diseño de Cubos
  - Herramienta de Análisis

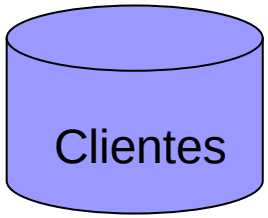
Planteo de la realidad

- Una empresa (*Tienda Celeste*), dedicada a la venta de diferentes tipos de ropa y calzado, que tiene sucursales en distintas ciudades, abarcando la mayoría de los departamentos.

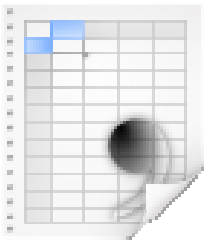
- Requerimientos:
  - Los gerentes desean analizar las *Ventas,* para poder definir la estrategia a futuro.
  - Interesa clasificarlas según los *Clientes*, el *Tiempo (fecha, mes y año)* y *tipo de Producto*.
  - Interesa evaluar las *ventas* por *ciudad* y *departamento de los Clientes*.

Fuentes

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



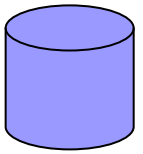
Productos (.xls)

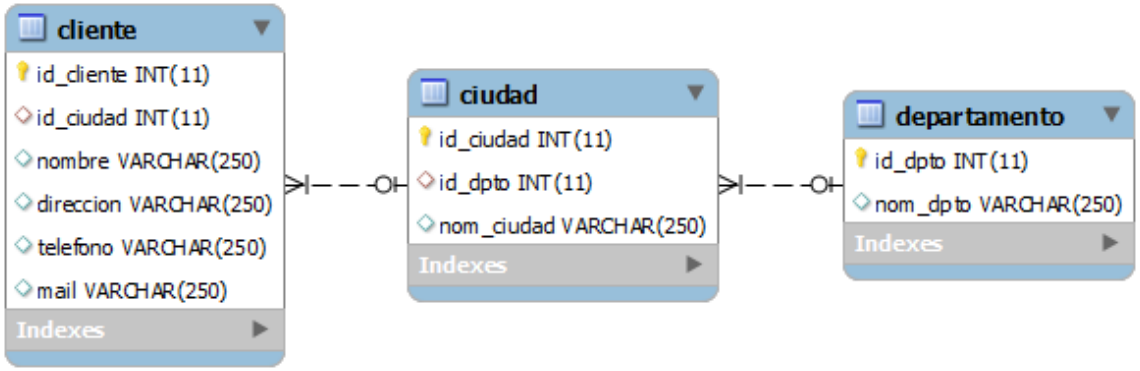


Ventas (.csv)

Fuentes

Clientes

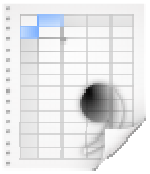




Productos

	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
	...	...	...	...

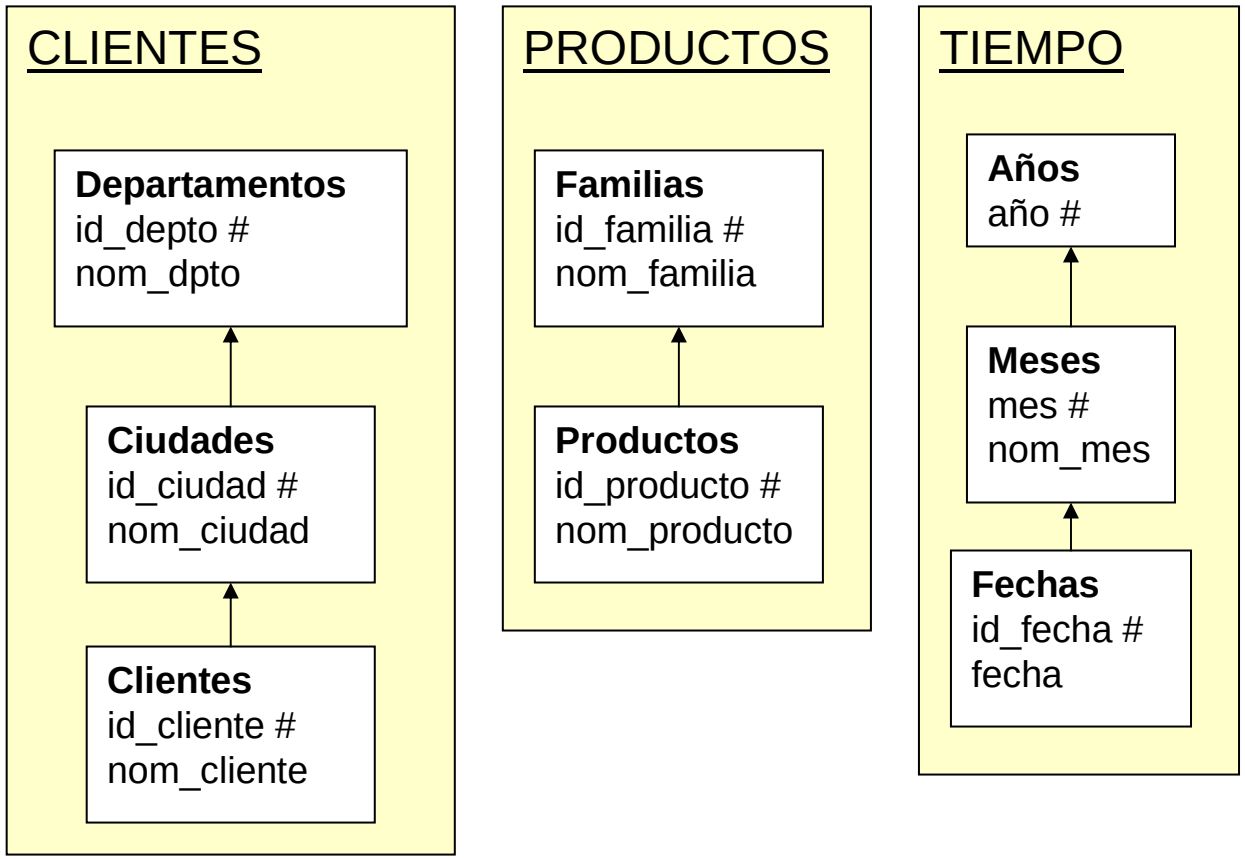
Ventas



	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
	...	...	...	...	...

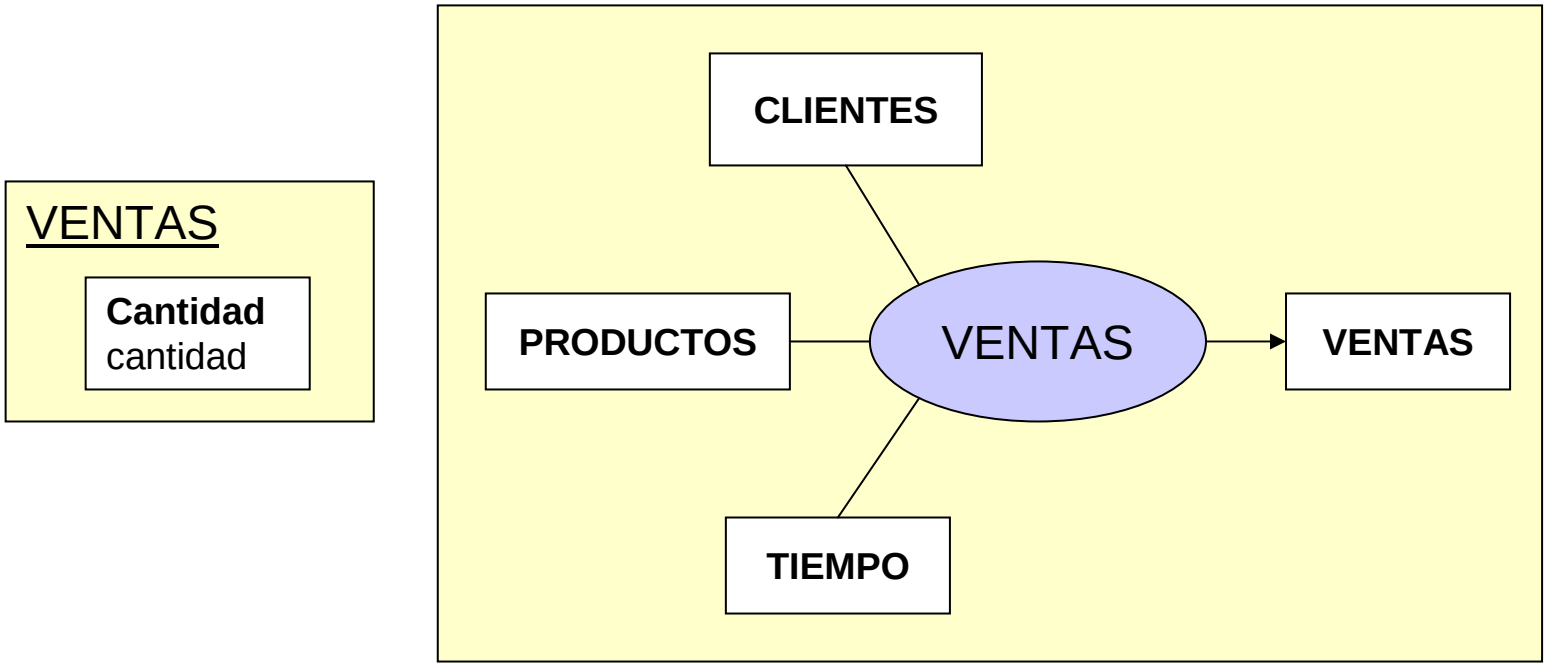
Diseño Conceptual

Dimensiones y Jerarquías



Diseño Conceptual

Relación Dimensional



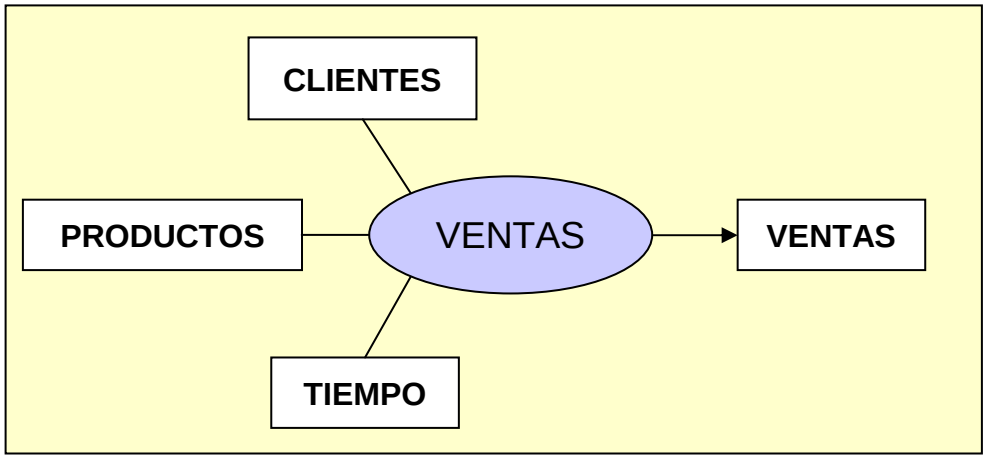
Diseño Conceptual

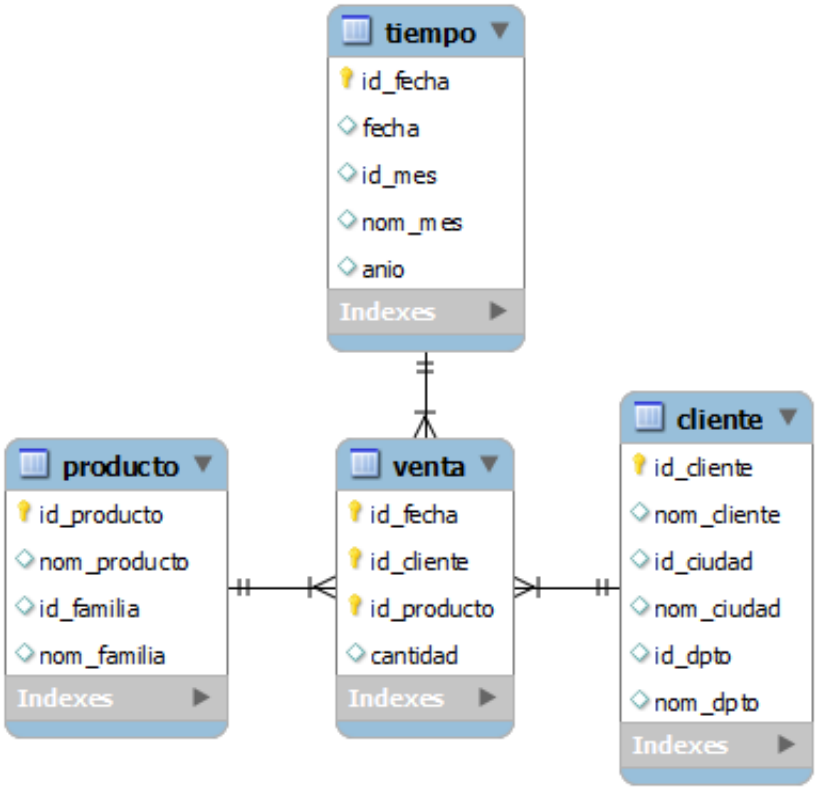
Estudio de Aditividad

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 9.6 Reference Manual

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.

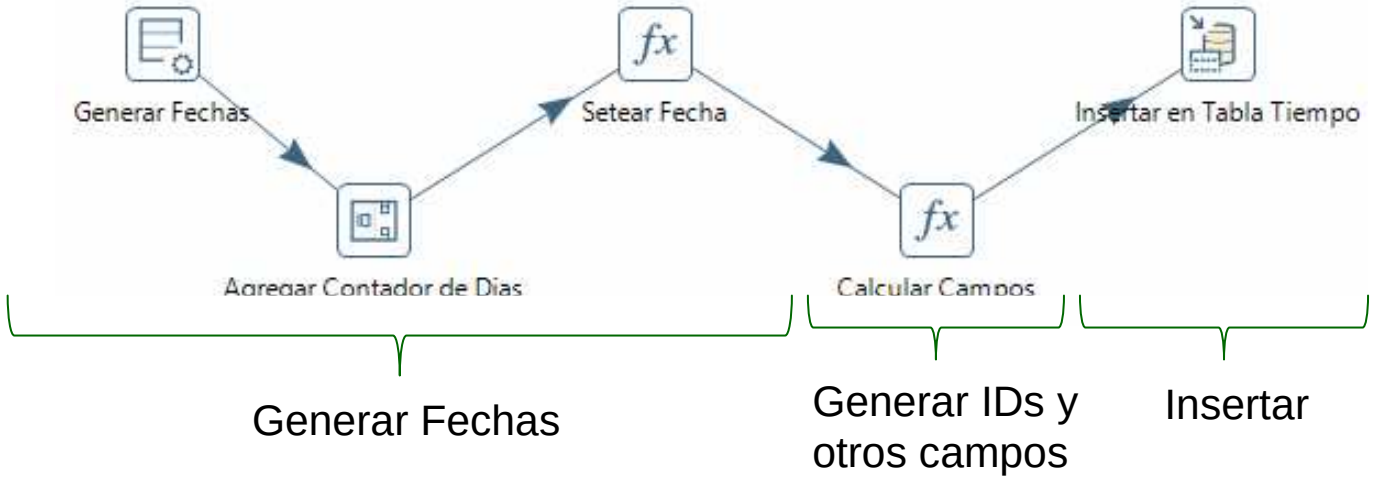
- Caso particular dimensión *Tiempo*. Claves subrogadas con formato `YYYYMMDD` permiten:
  - a partir de una fecha generar el identificador y viceversa
  - preservan la relación de orden

Diseño Físico

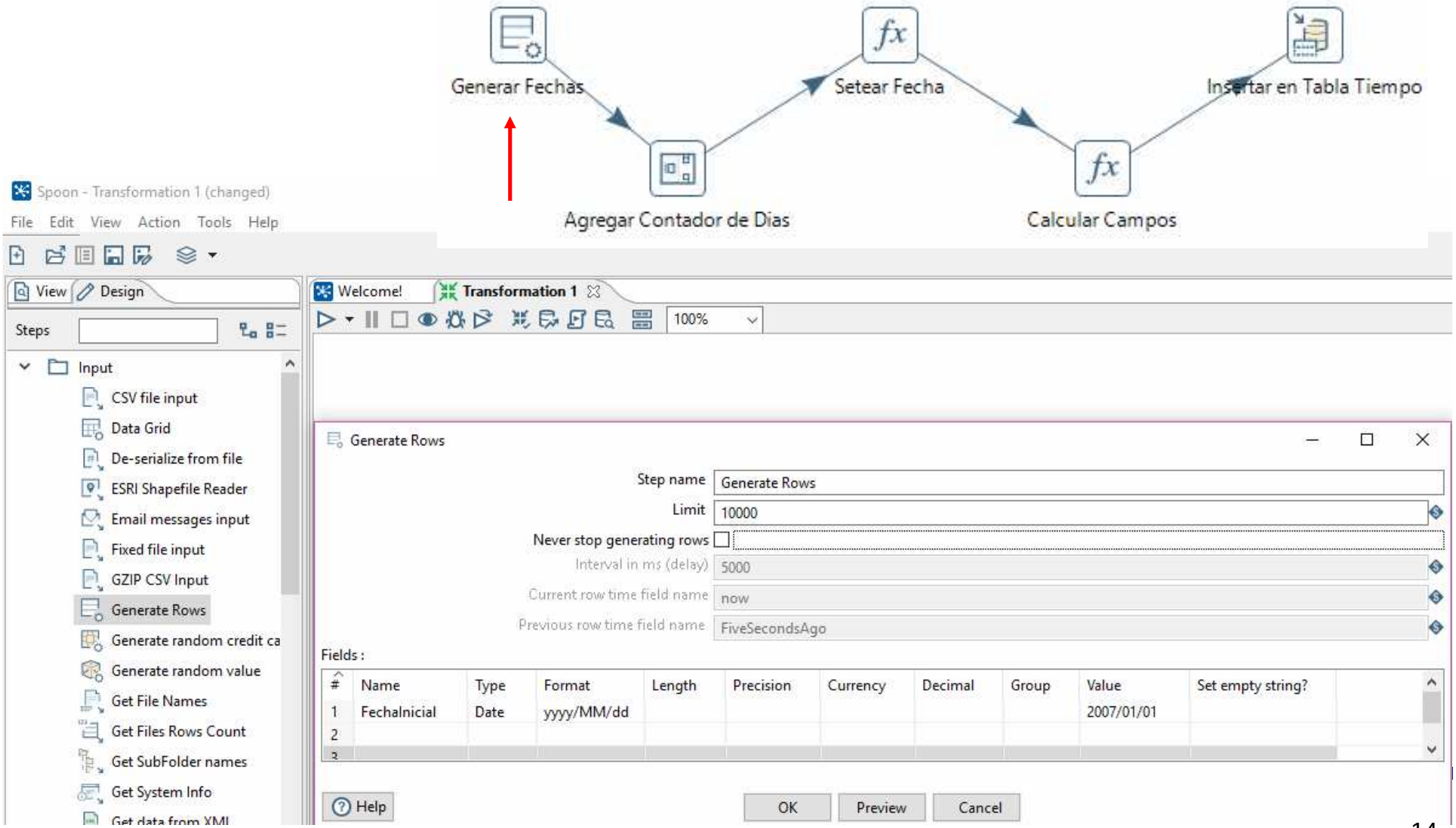
- Seleccionar un conjunto de índices adecuado, que permita realizar joins de forma eficiente: B-TREE, HASH, etc.
- Estudiar características particulares del DBMS para aprovecharlo al máximo. Por ejemplo en PostgreSQL:
  - Table partitioning: Divide tablas muy grandes en pequeñas tablas (https://www.postgresql.org/docs/9.6/static/ddl-partitioning.html)
  - VACUUM: Recupera el almacenamiento ocupado por «tuplas muertas» (https://www.postgresql.org/docs/9.6/static/sql-vacuum.html)
  - ANALIZE: Recopila estadísticos de la BD (https://www.postgresql.org/docs/9.6/static/sql-analyze.html)
  - REINDEX: Reconstruye índices (https://www.postgresql.org/docs/9.6/static/sql-reindex.html)

Carga de Dimensiones

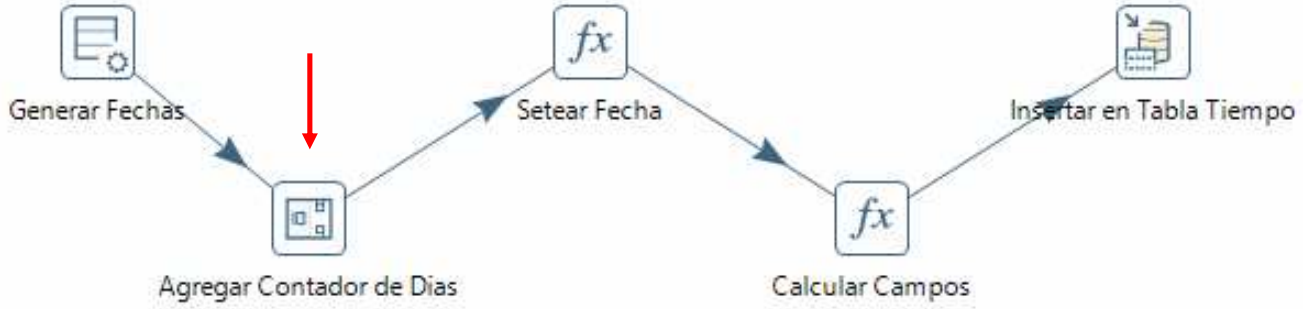
- Dimensión *Tiempo*:
  - Generar las fechas
  - Para cada fecha generar los identificadores y los restantes campos de la tabla.
  - Insertar en la tabla *Tiempo* del DW.

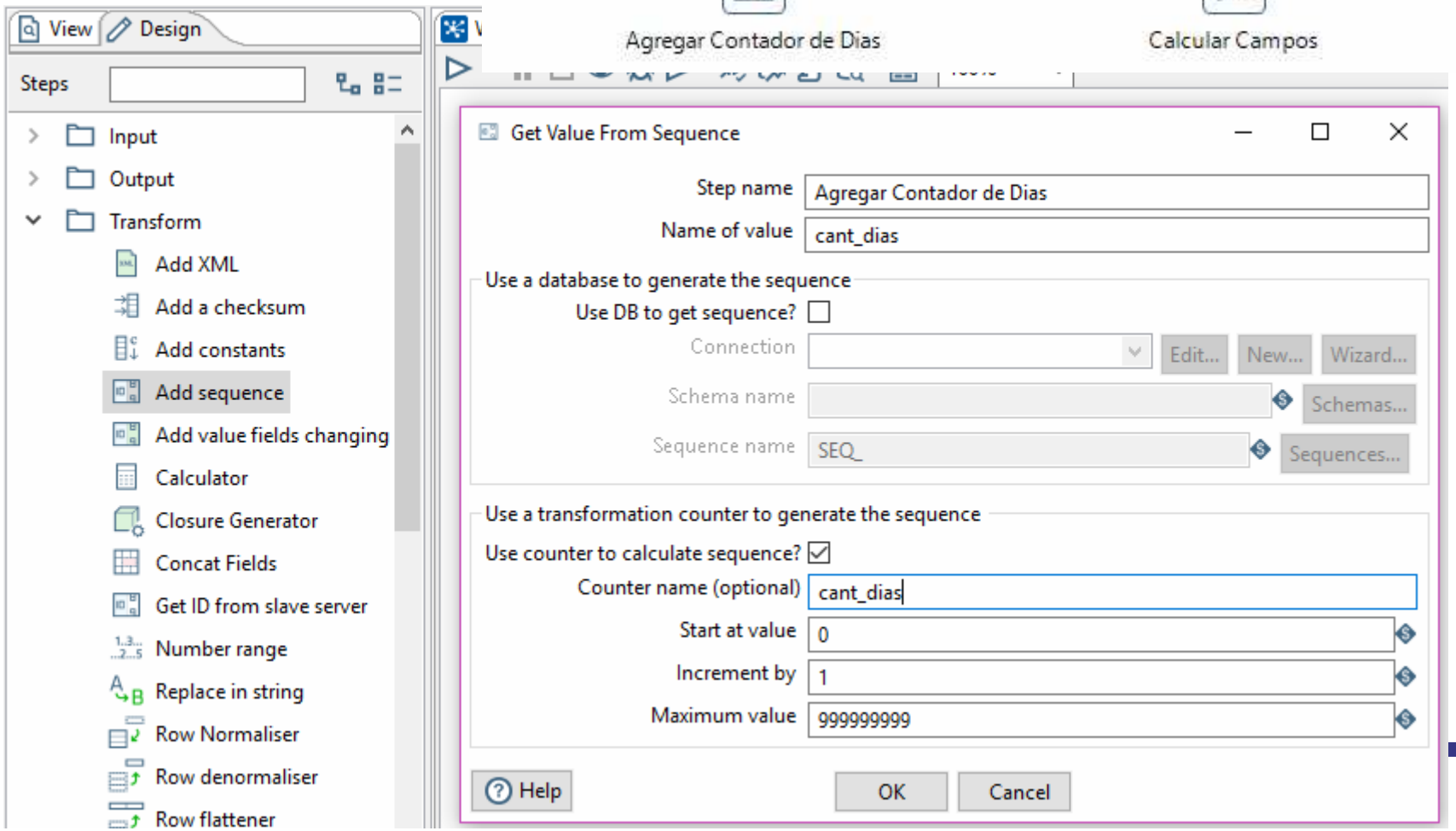


Carga de Dimensión Tiempo

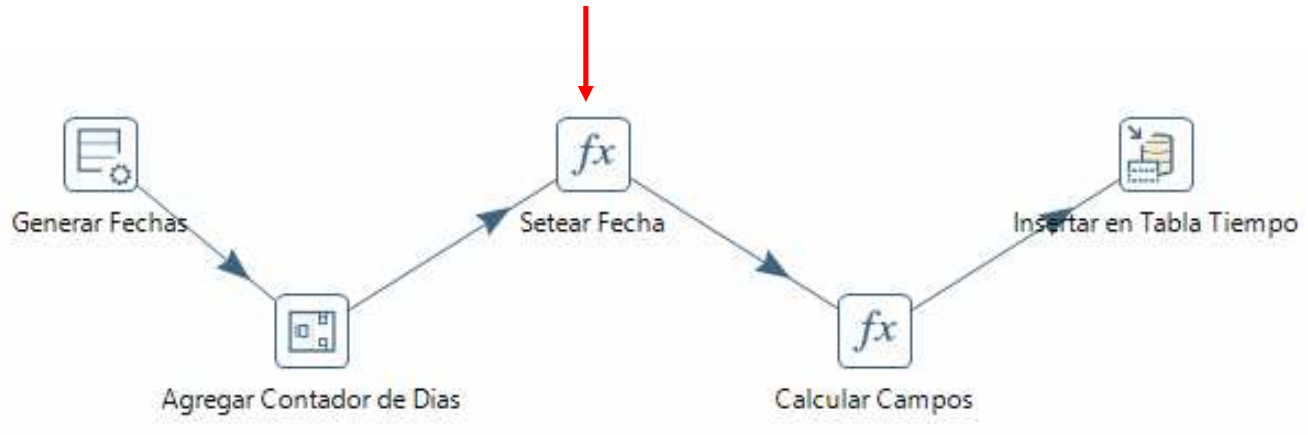


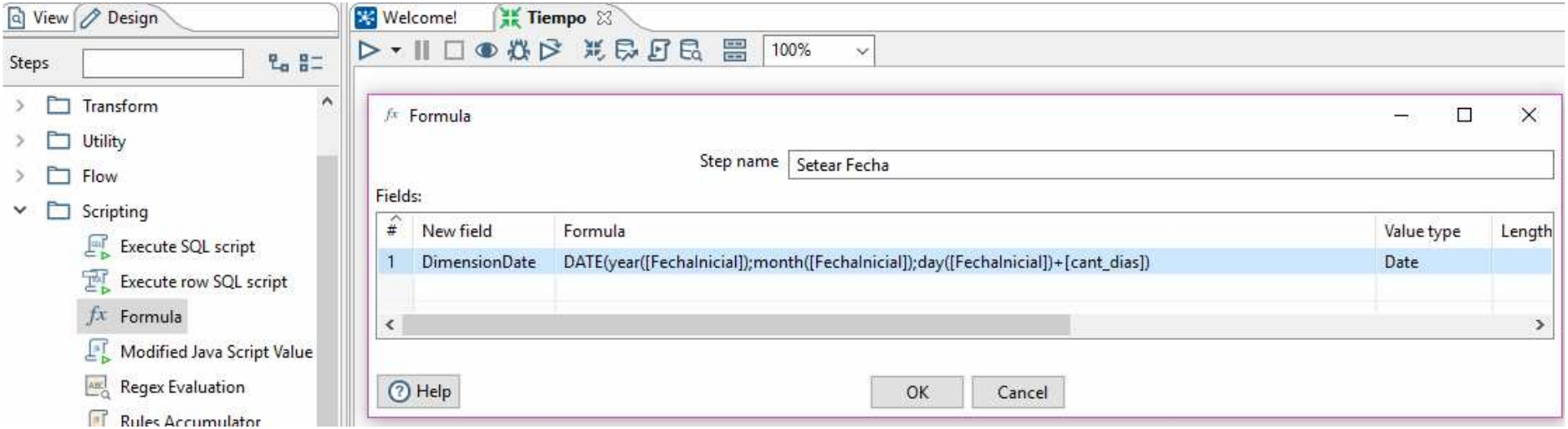
Carga de Dimensión Tiempo



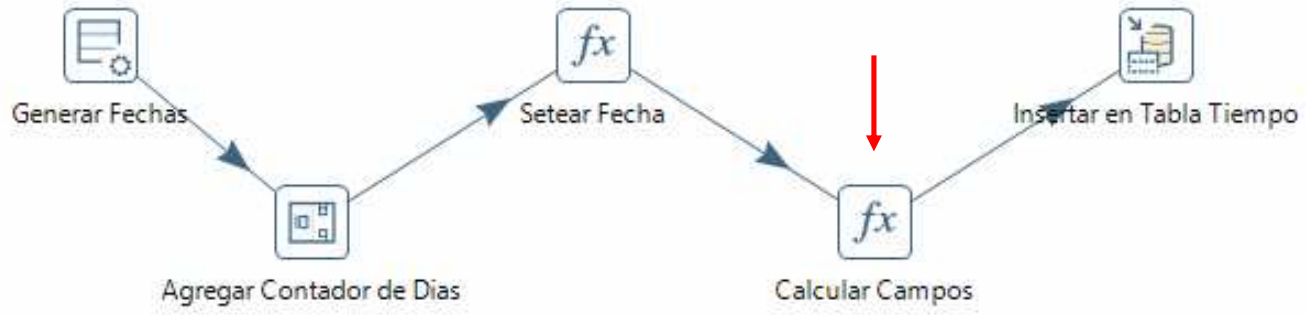


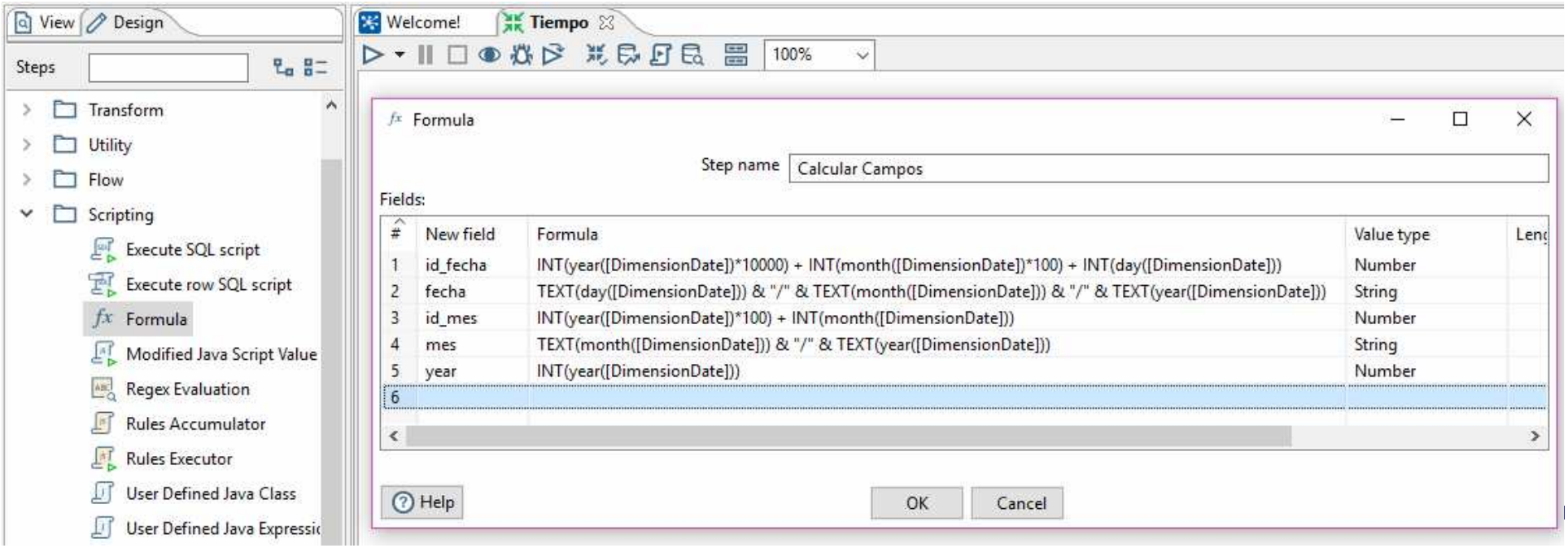
Carga de Dimensión Tiempo



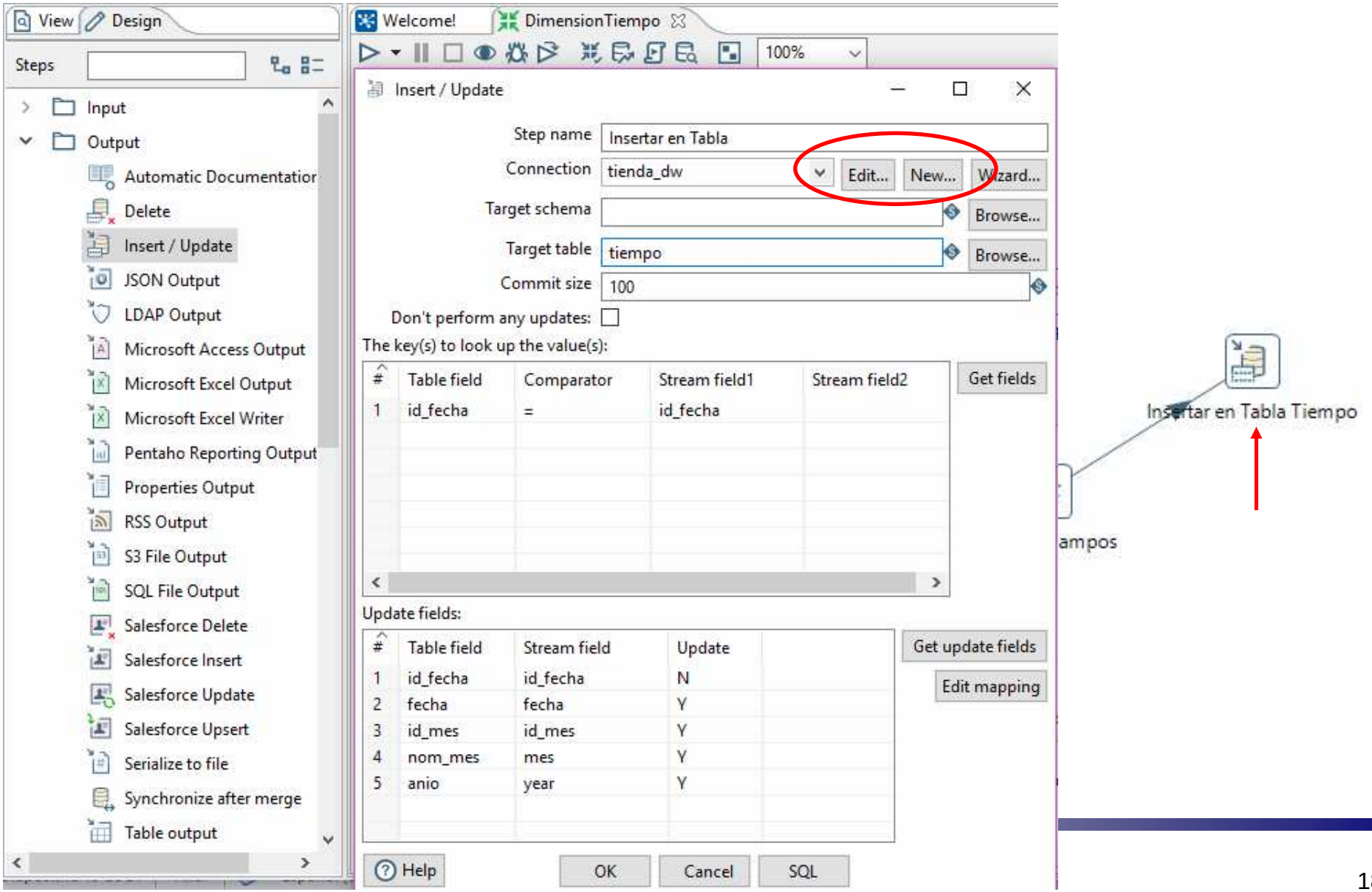


Carga de Dimensión Tiempo

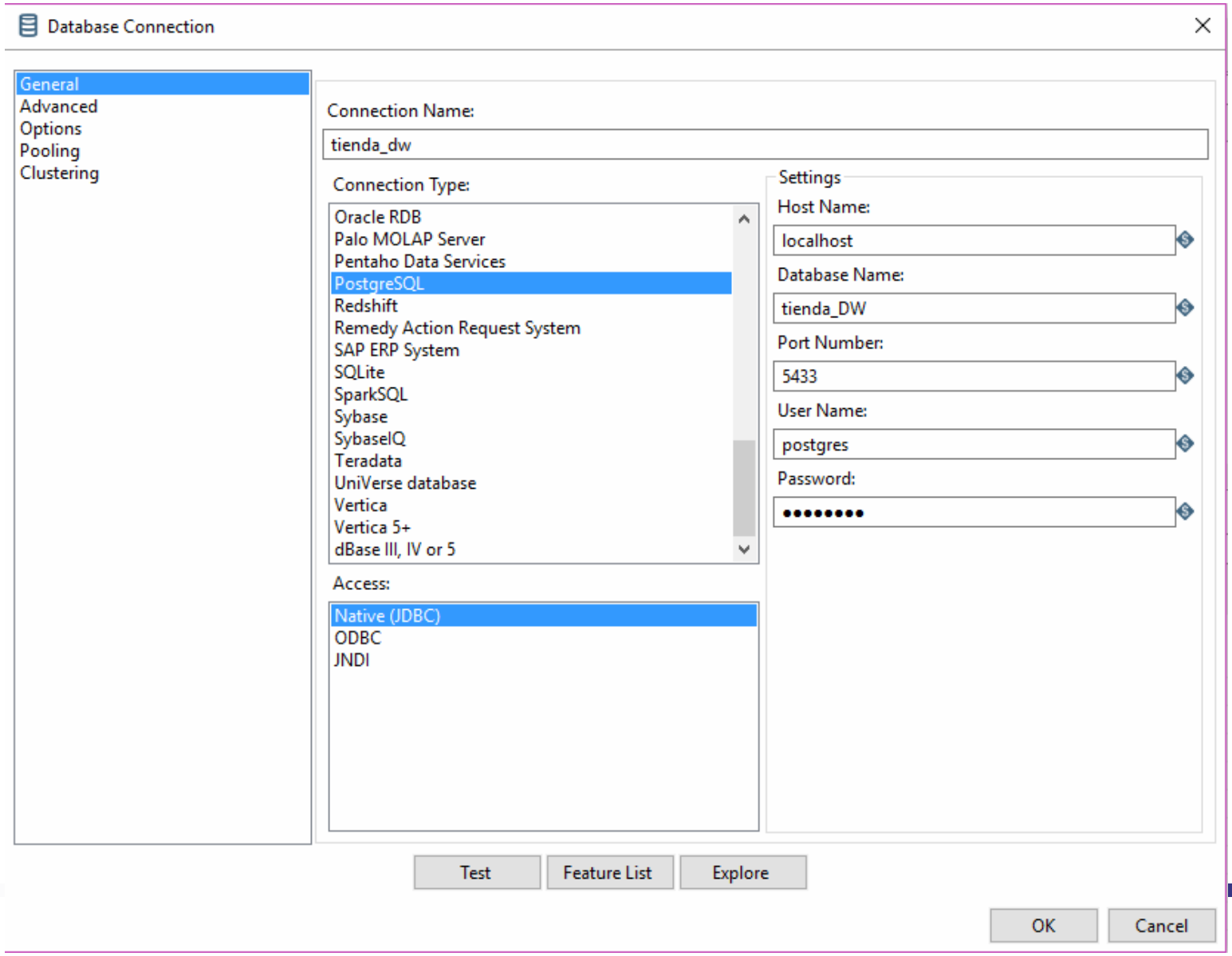




Carga de Dimensión Tiempo

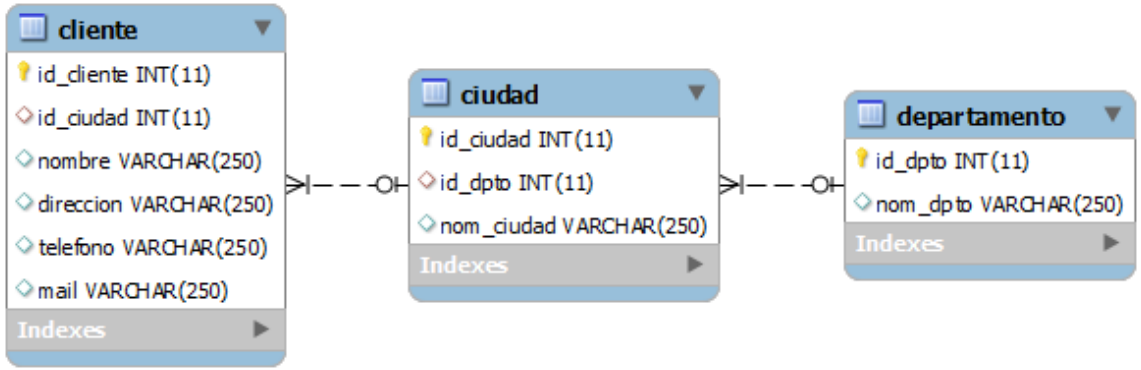


Carga de Dimensión Tiempo

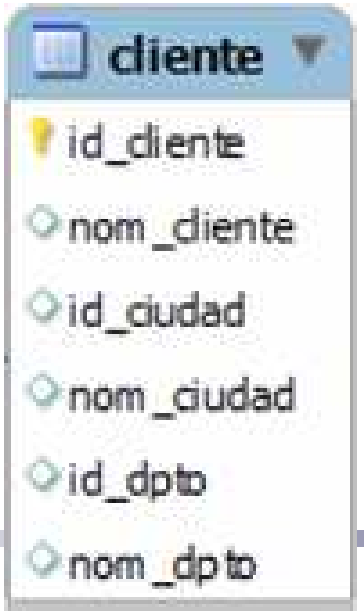


Carga de Dimensiones

- Cliente
  - BD Fuente (BD):



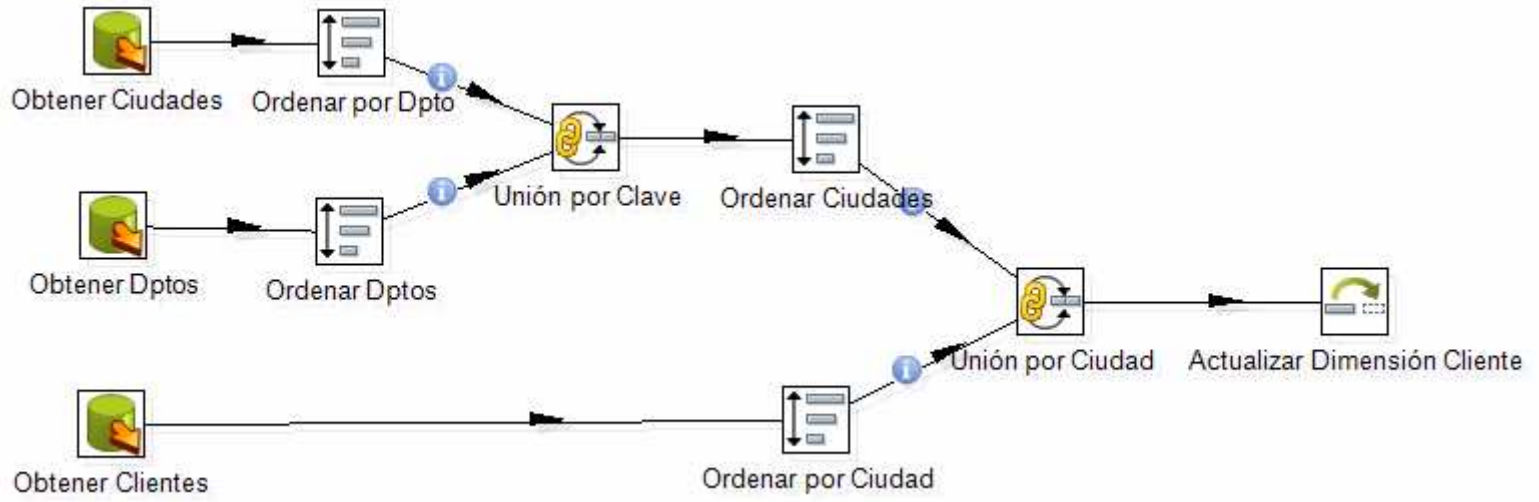
  - Tabla en el DW:



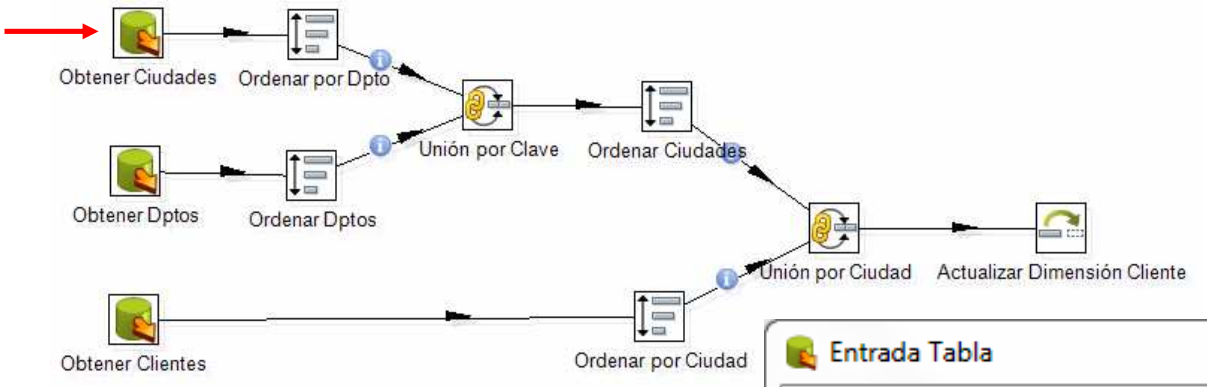
Carga de Dimensión Cliente

- 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

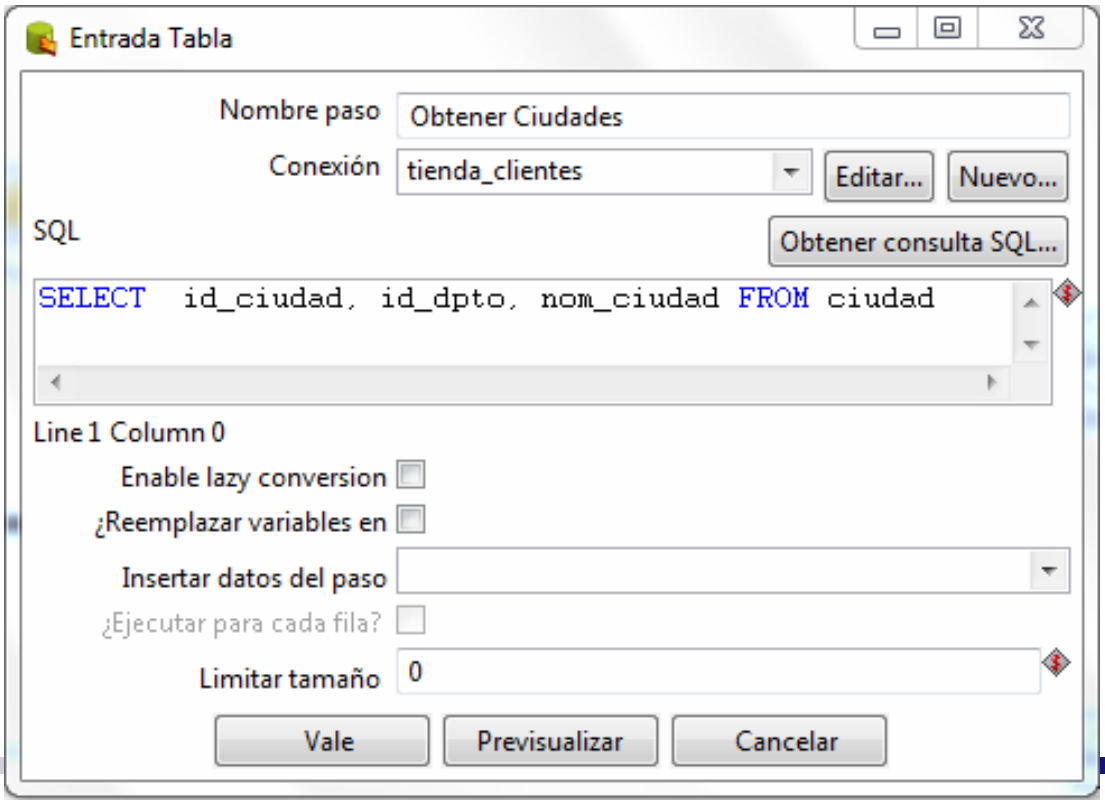


Carga de Dimensión Cliente

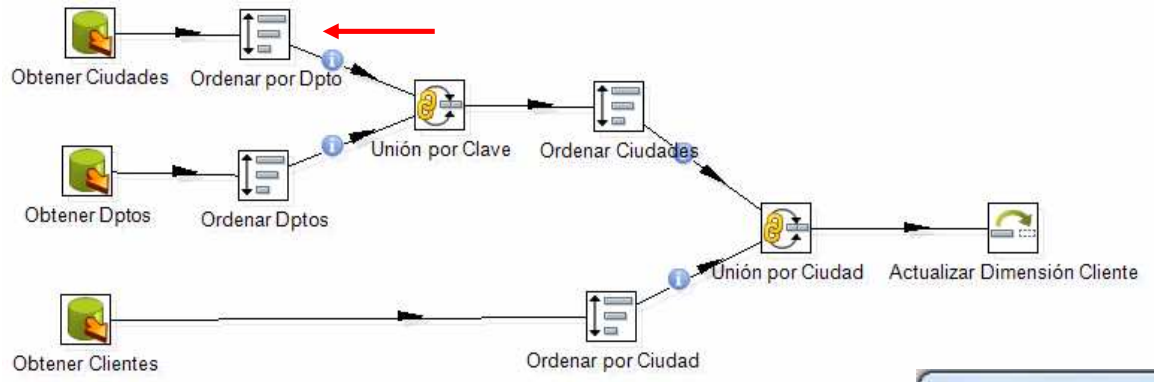


Cliente: *Step*

«Obtener ciudades»

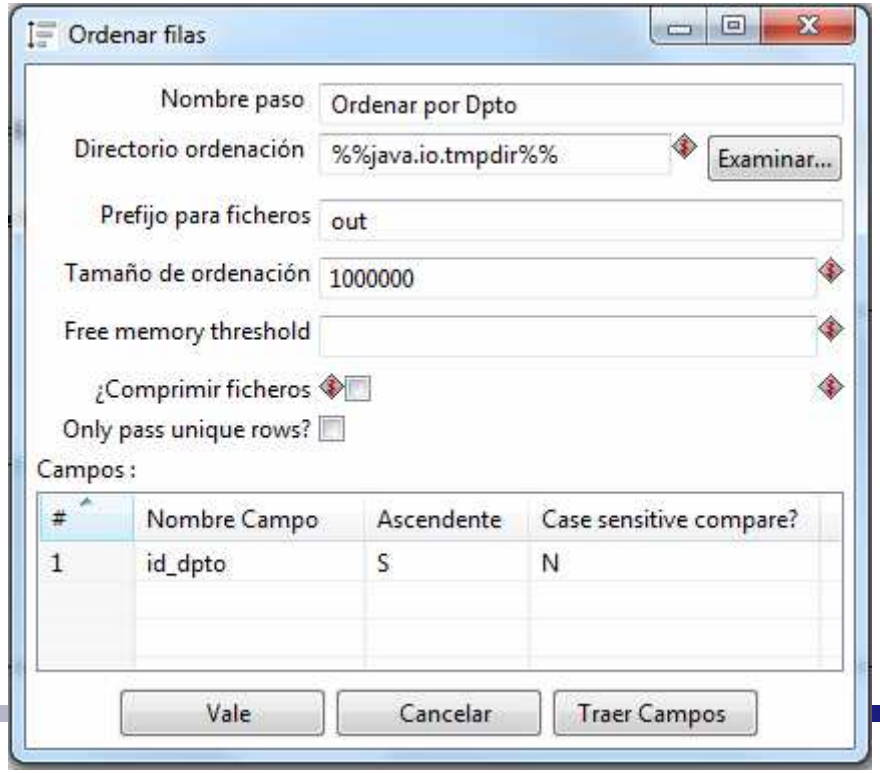


Carga de Dimensión Cliente

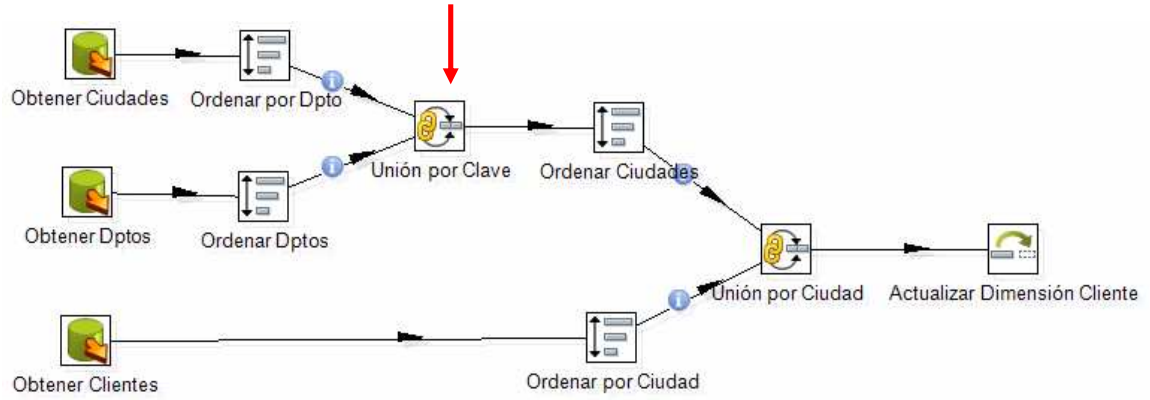


Cliente: *Step*

«Ordenar por Dpto.»

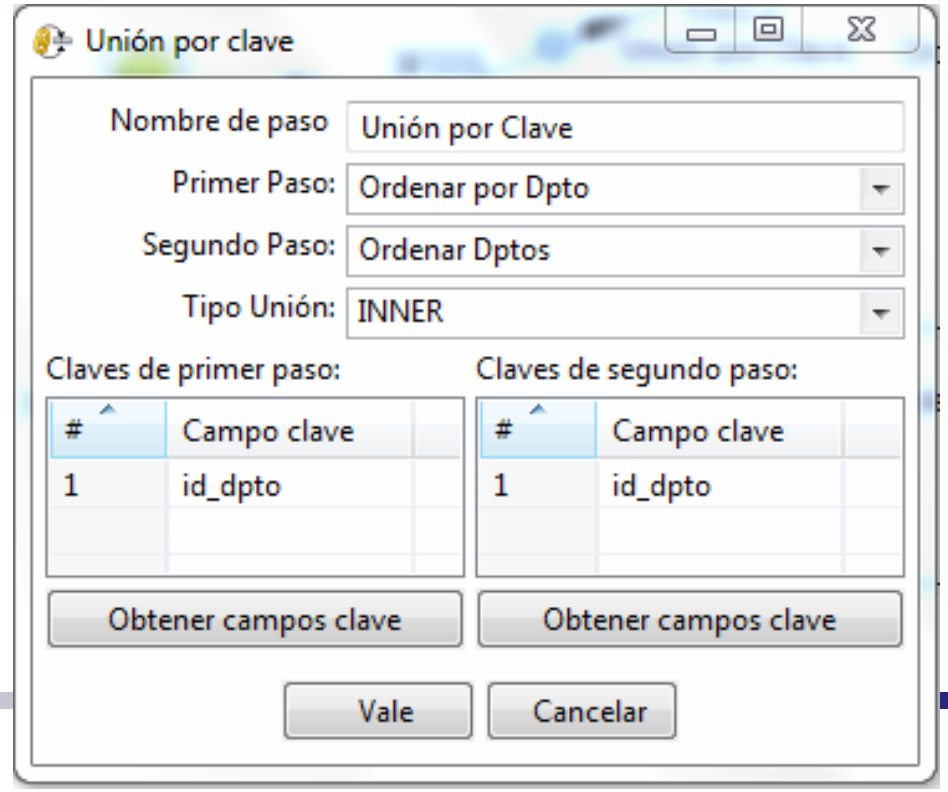


Carga de Dimensión Cliente

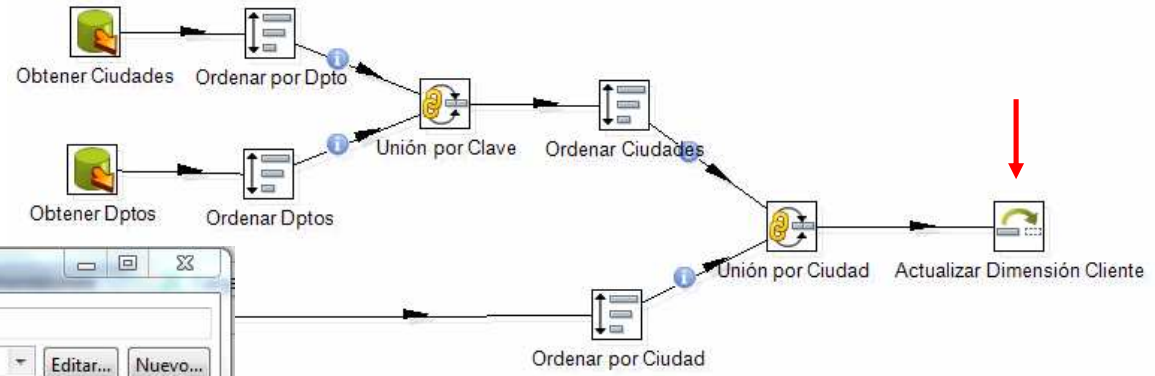


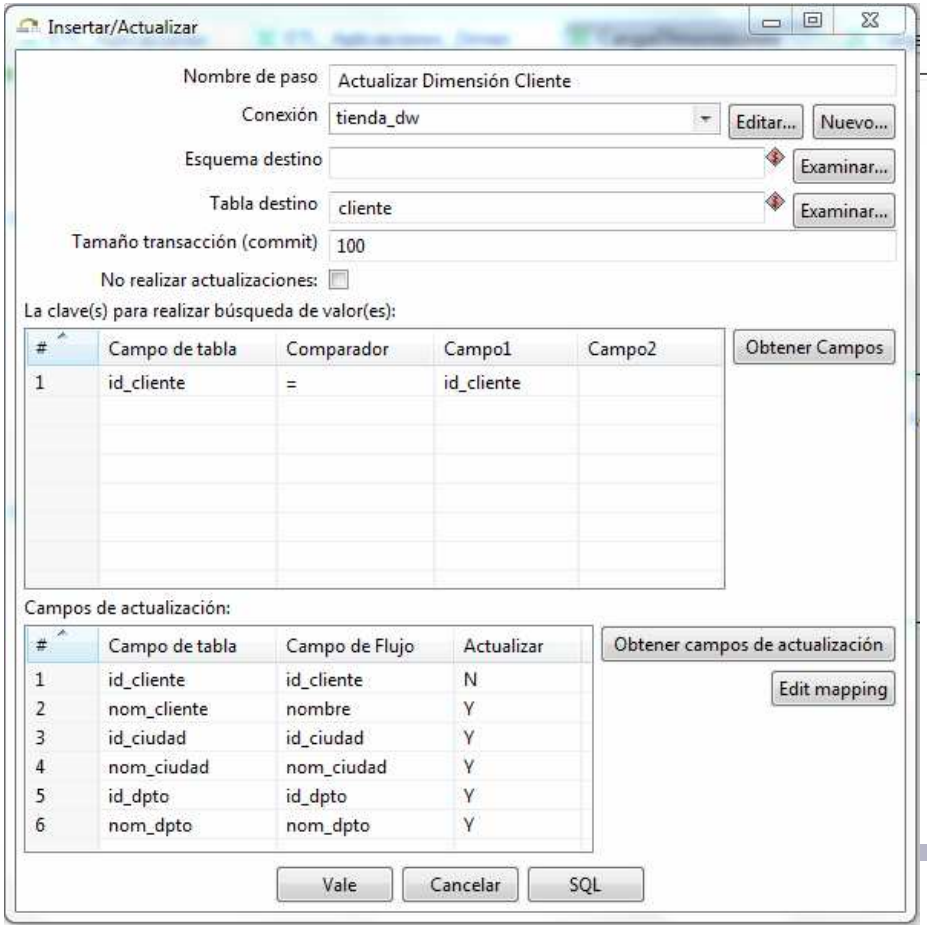
Cliente: *Step*

«Unión por Clave»



Carga de Dimensión Cliente





Cliente: *Step*

«Actualizar dimensión Cliente»

Carga de Dimensiones

- 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:

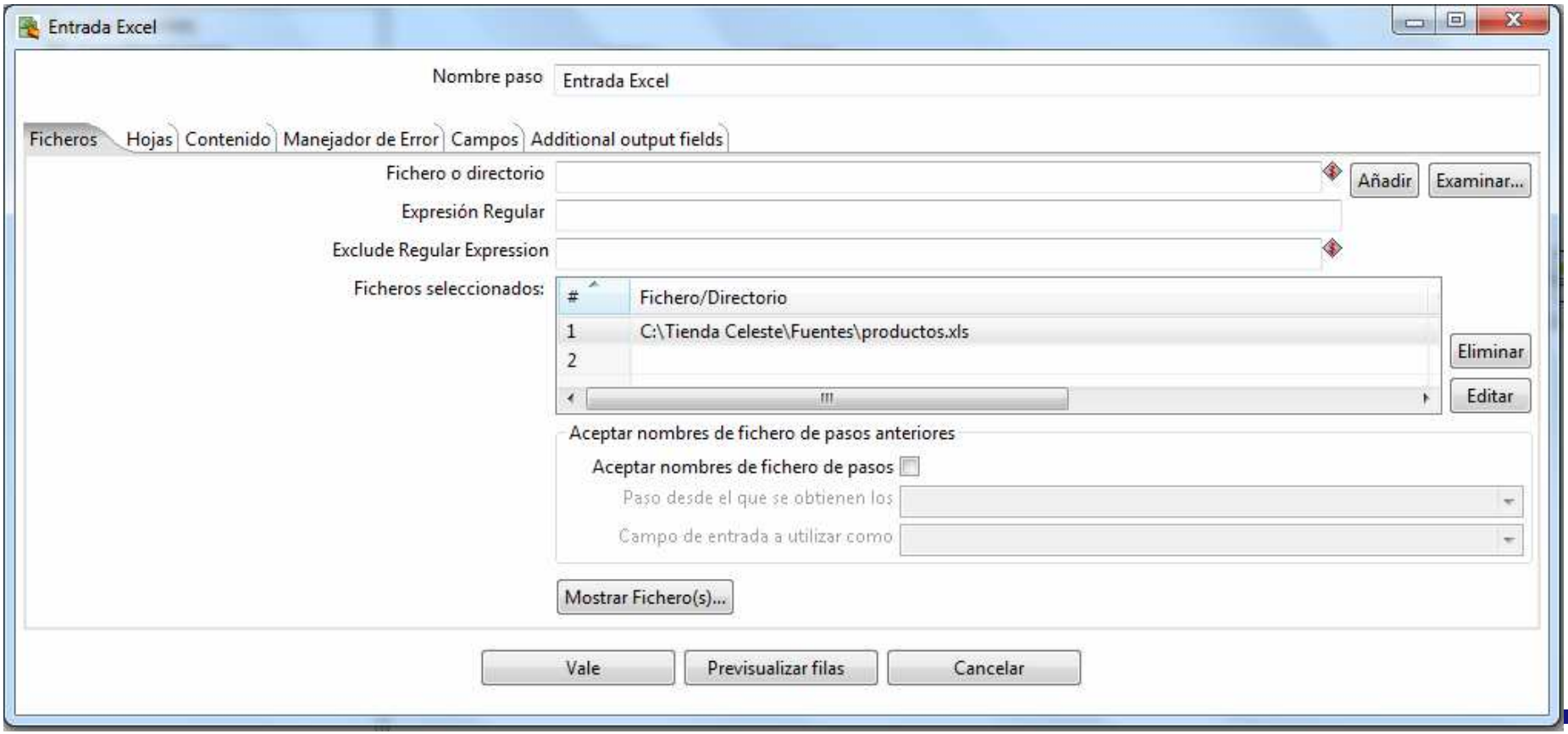
Carga de Dimensión Producto

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

- Producto: Transformación

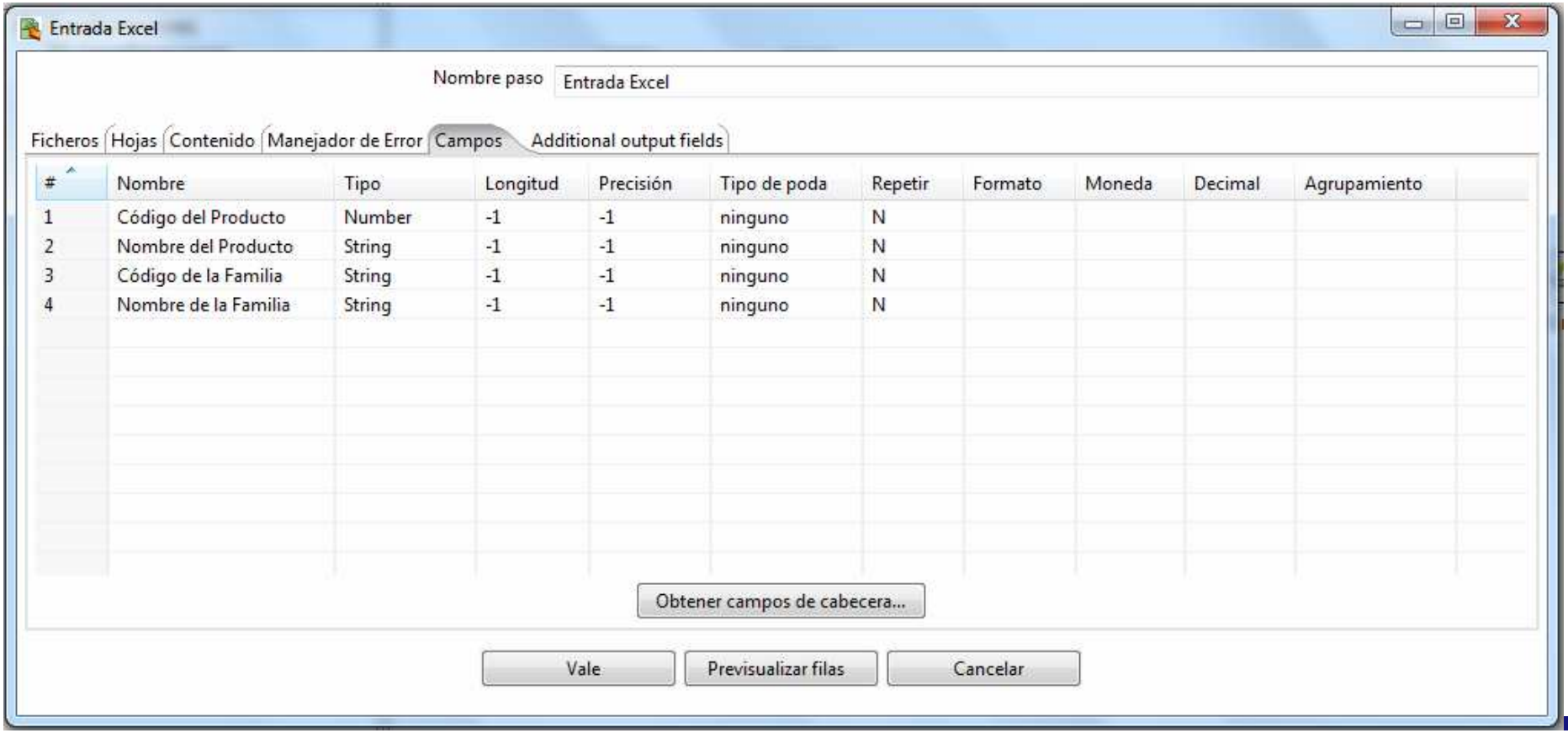
Carga de Dimensión Producto

Producto: *Step «Entrada Excel»*



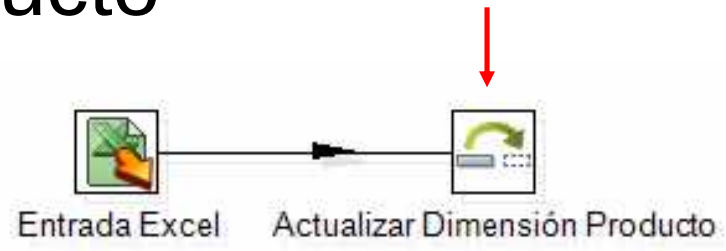
Carga de Dimensión Producto

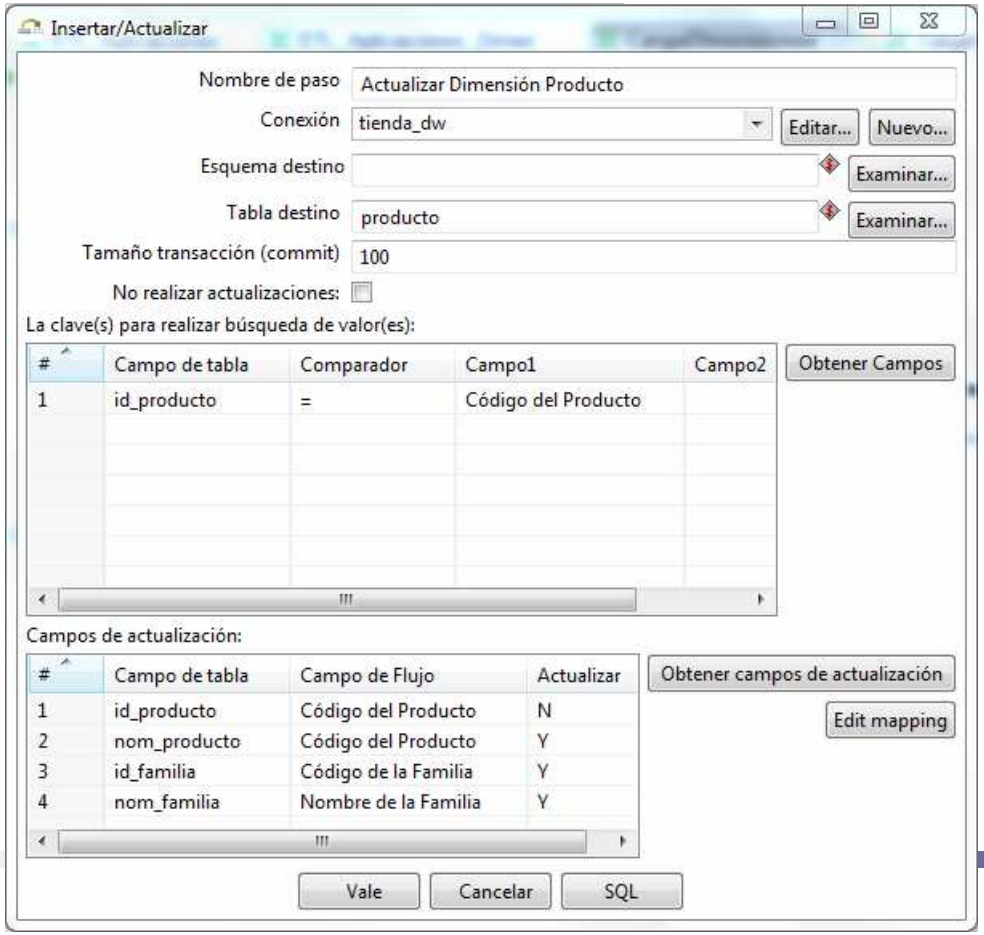
Producto: *Step «Entrada Excel»*



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:

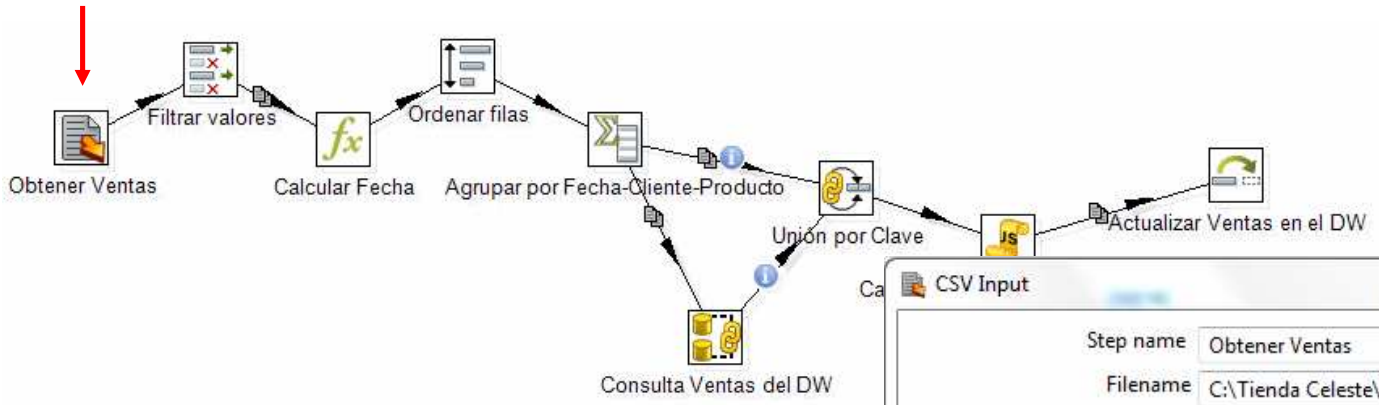
Carga de Hechos

- 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

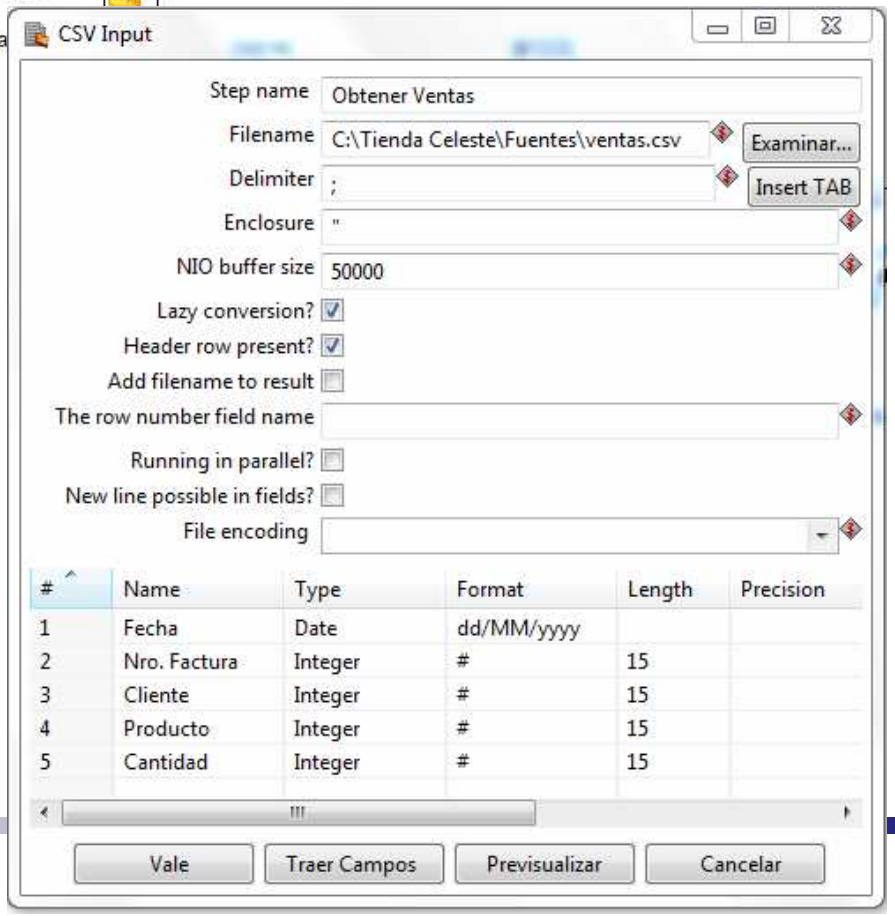
- Ventas: Transformación

Carga de Hechos



Ventas: *Step*

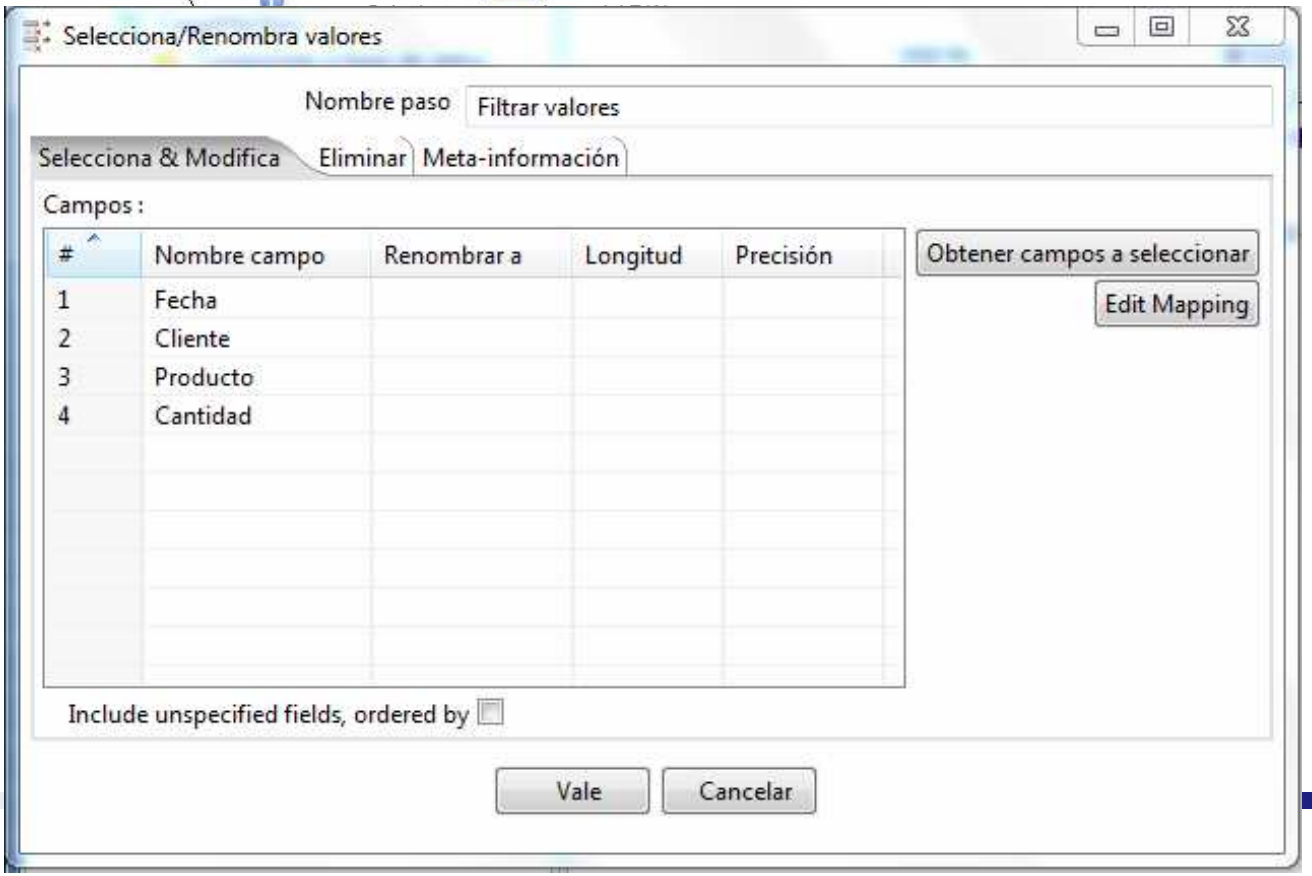
«Obtener Ventas»



Carga de Hechos

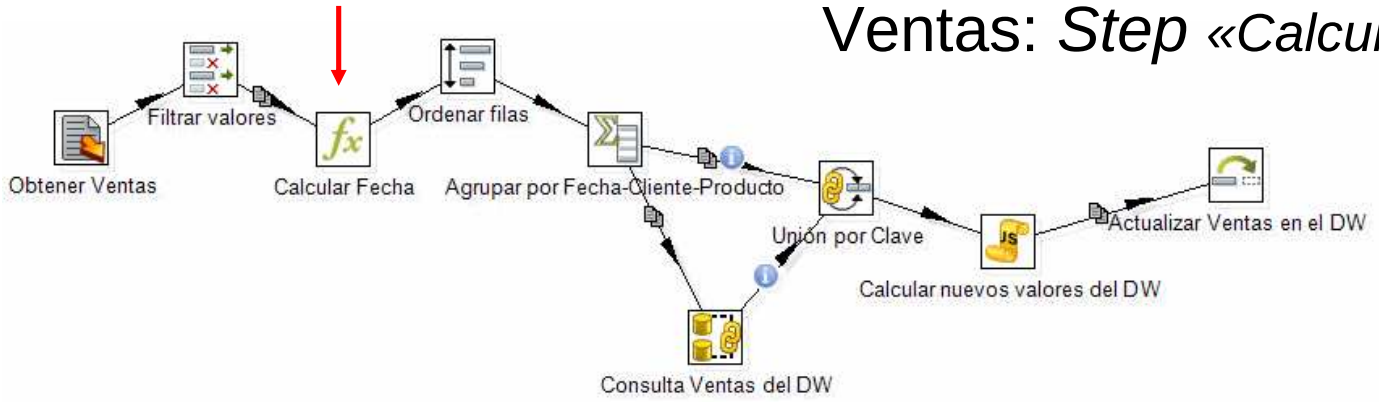
Ventas: *Step*

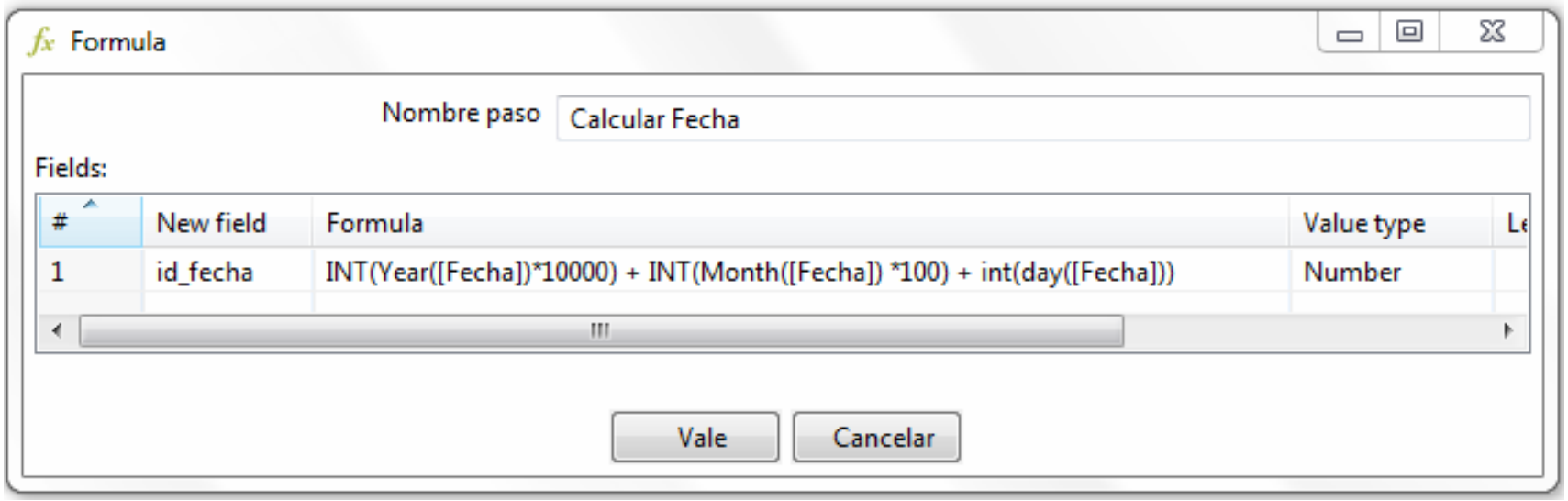
«Filtrar valores»



Carga de Hechos

Ventas: *Step «Calcular Fecha»*



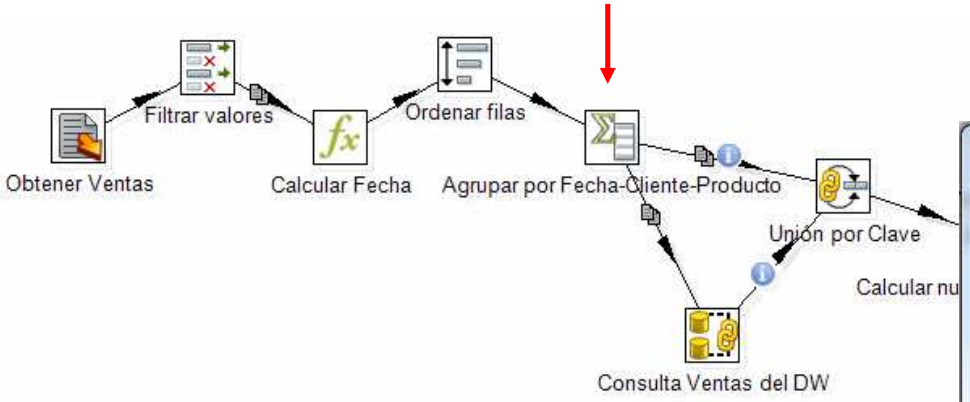


Carga de Hechos

Ventas: *Step*

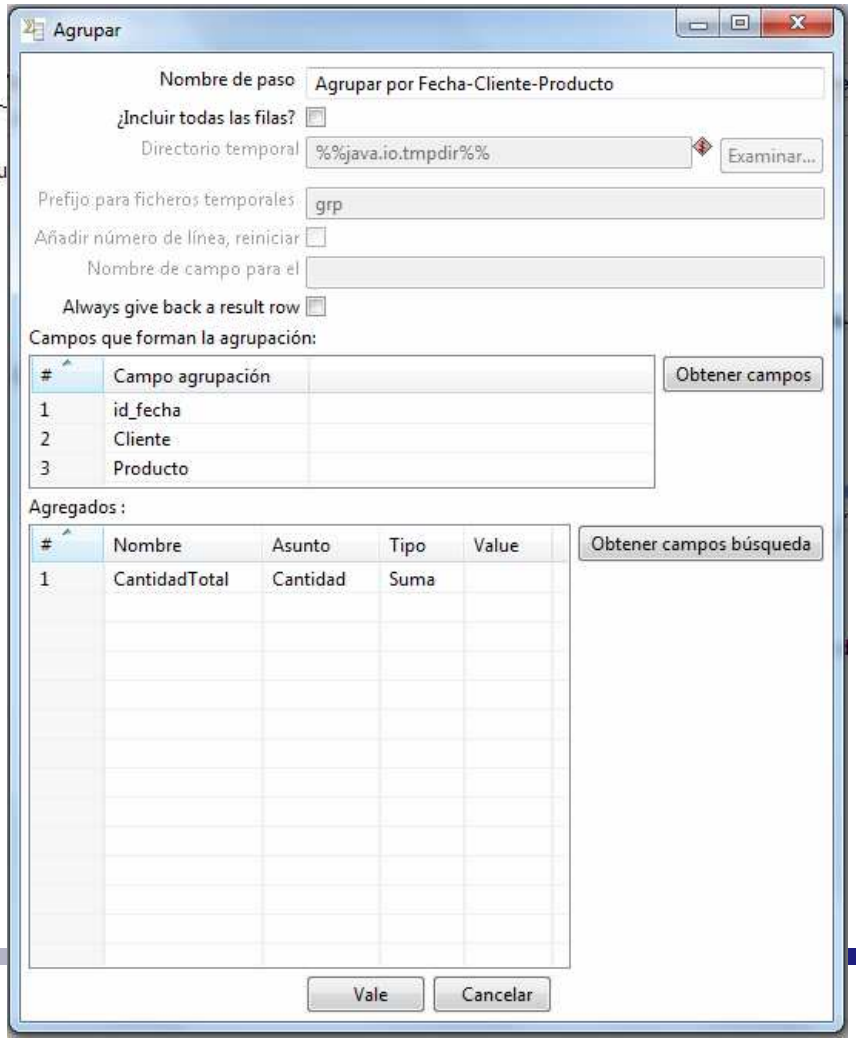
«Ordenar filas»

Carga de Hechos



Ventas: *Step*

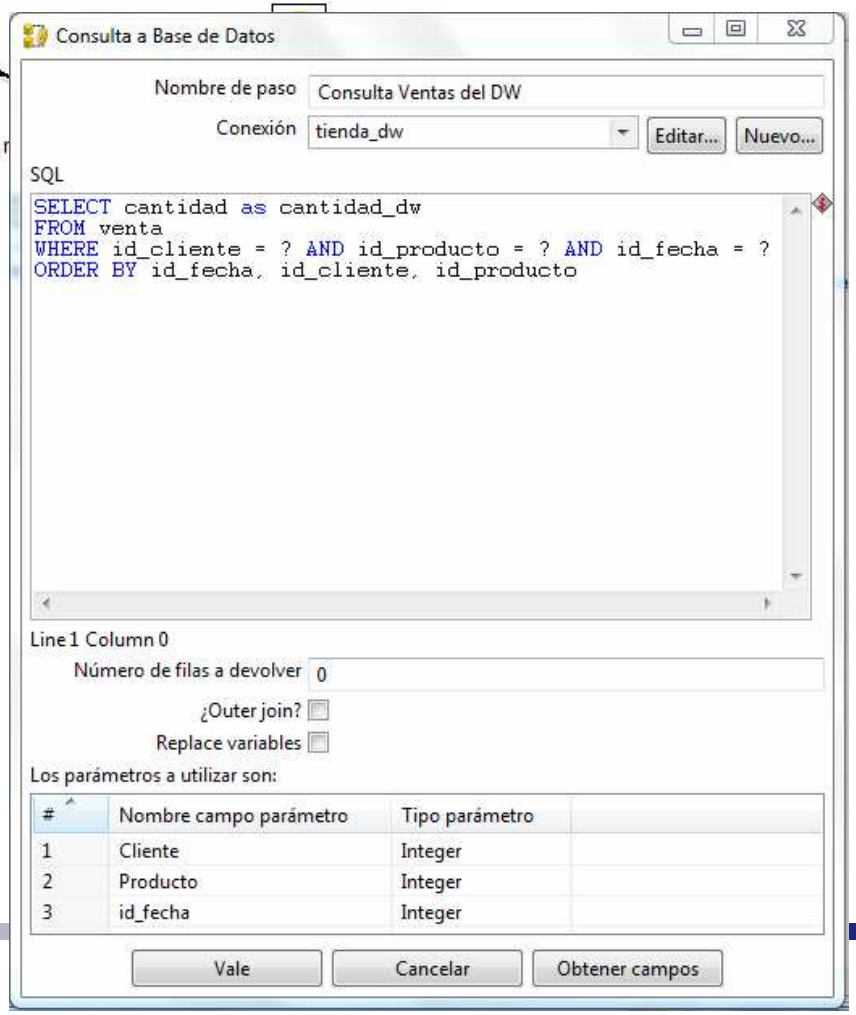
«Agrupar por Fecha-Cliente-Producto»



Carga de Hechos

Ventas: *Step*

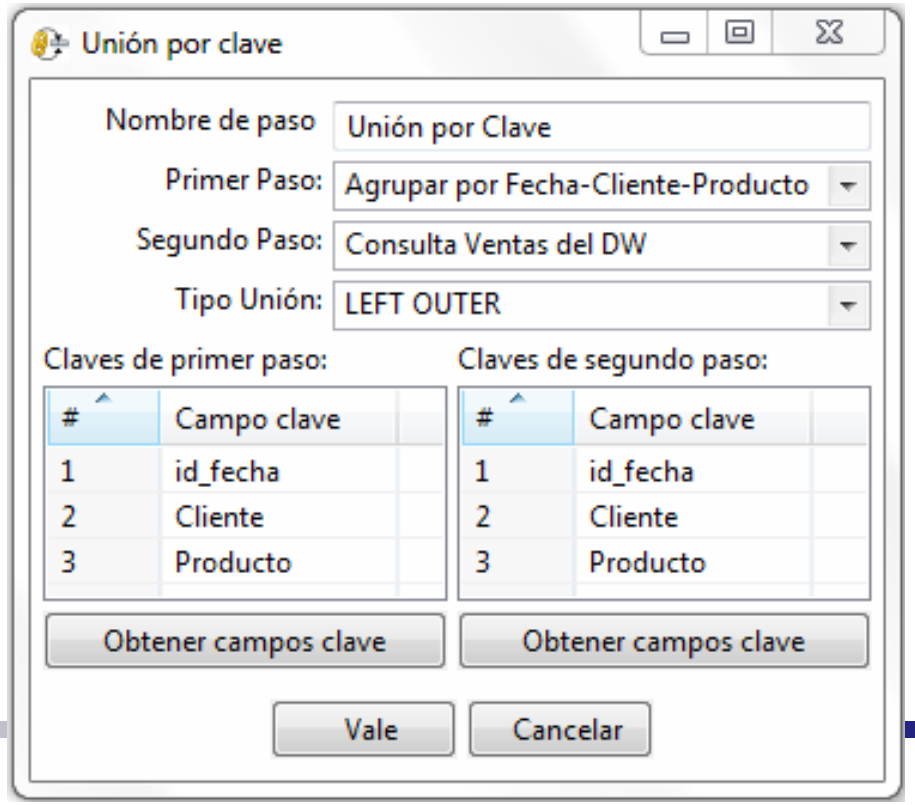
«Consultas Ventas del DW»



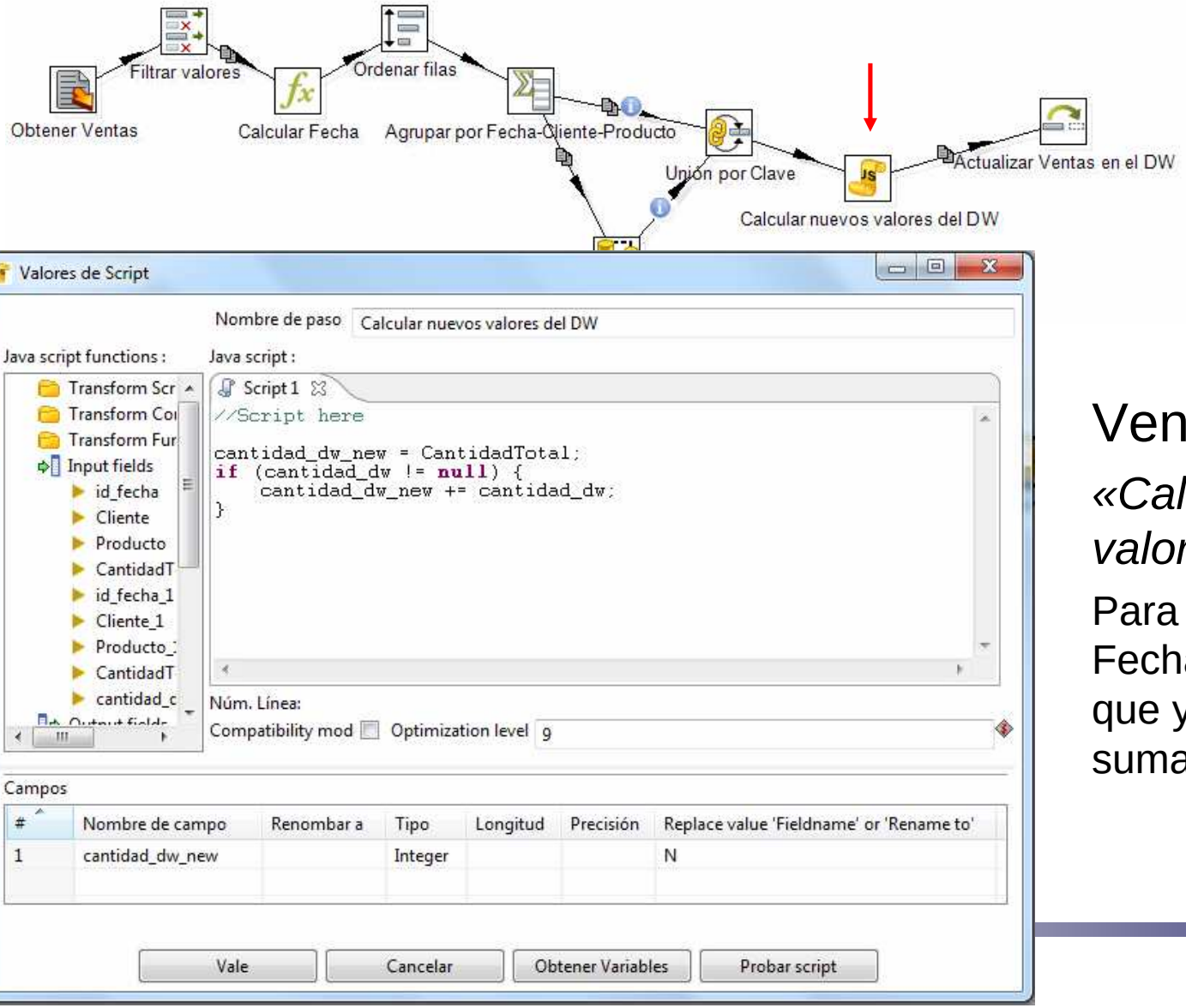
Carga de Hechos

Ventas: *Step*

«Unión por Clave»



Carga de Hechos

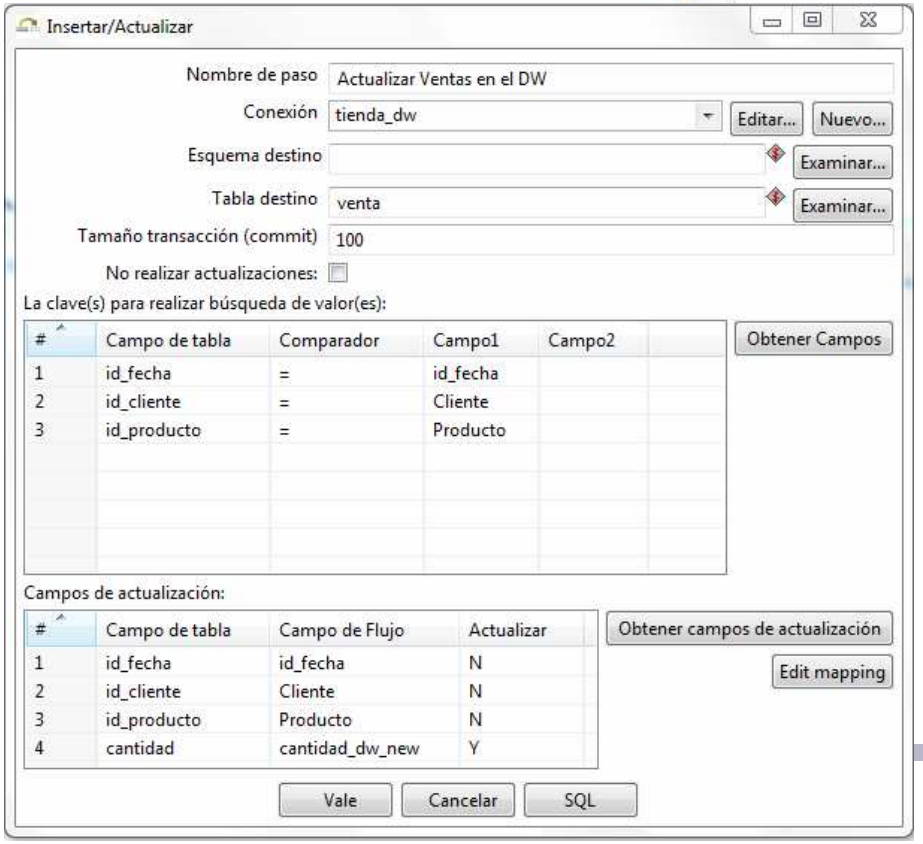


Ventas: *Step*

«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

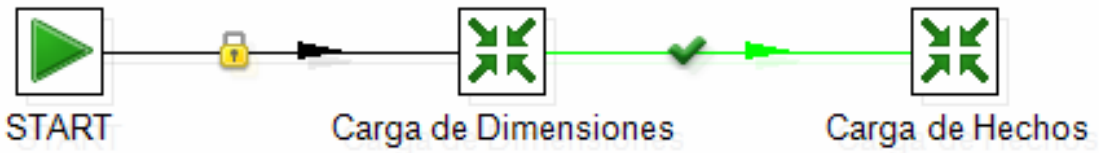
- En general, la dimensión *Tiempo* podría ser cargada solo la primera vez, eventualmente cada mucho tiempo.

- 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)

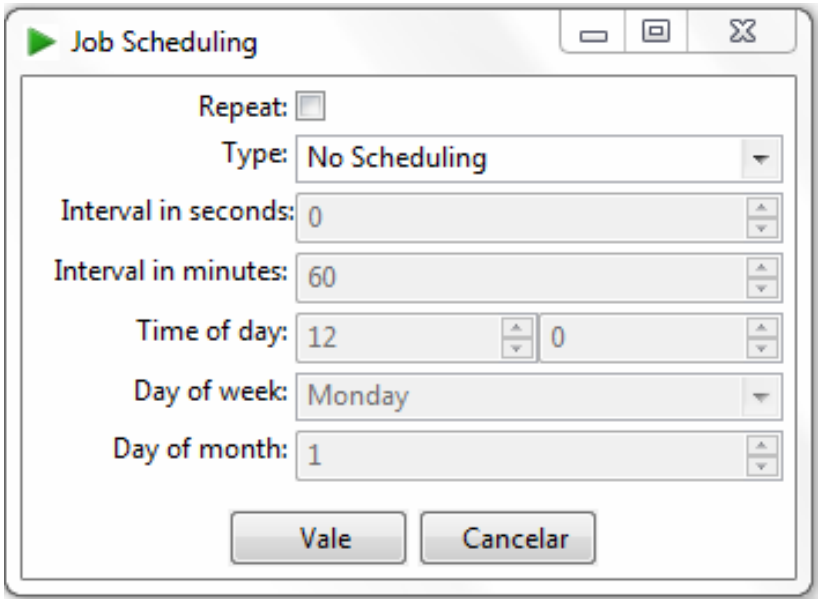
- Primera Carga:
  - Cargar todas las Dimensiones
  - Cargar los Hechos

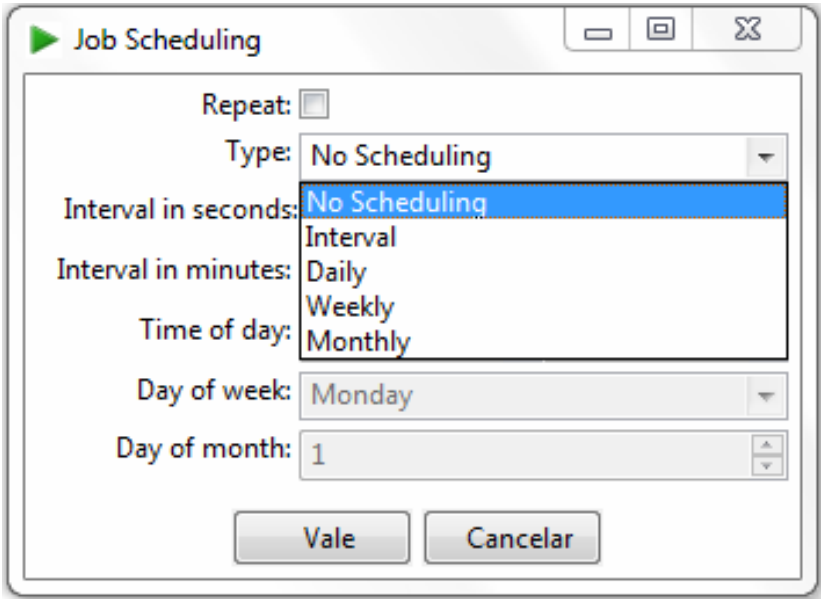
- Actualización del DW:
  - Cargar las Dimensiones “dinámicas” *Clientes y Productos*
  - Cargar los Hechos *Ventas*



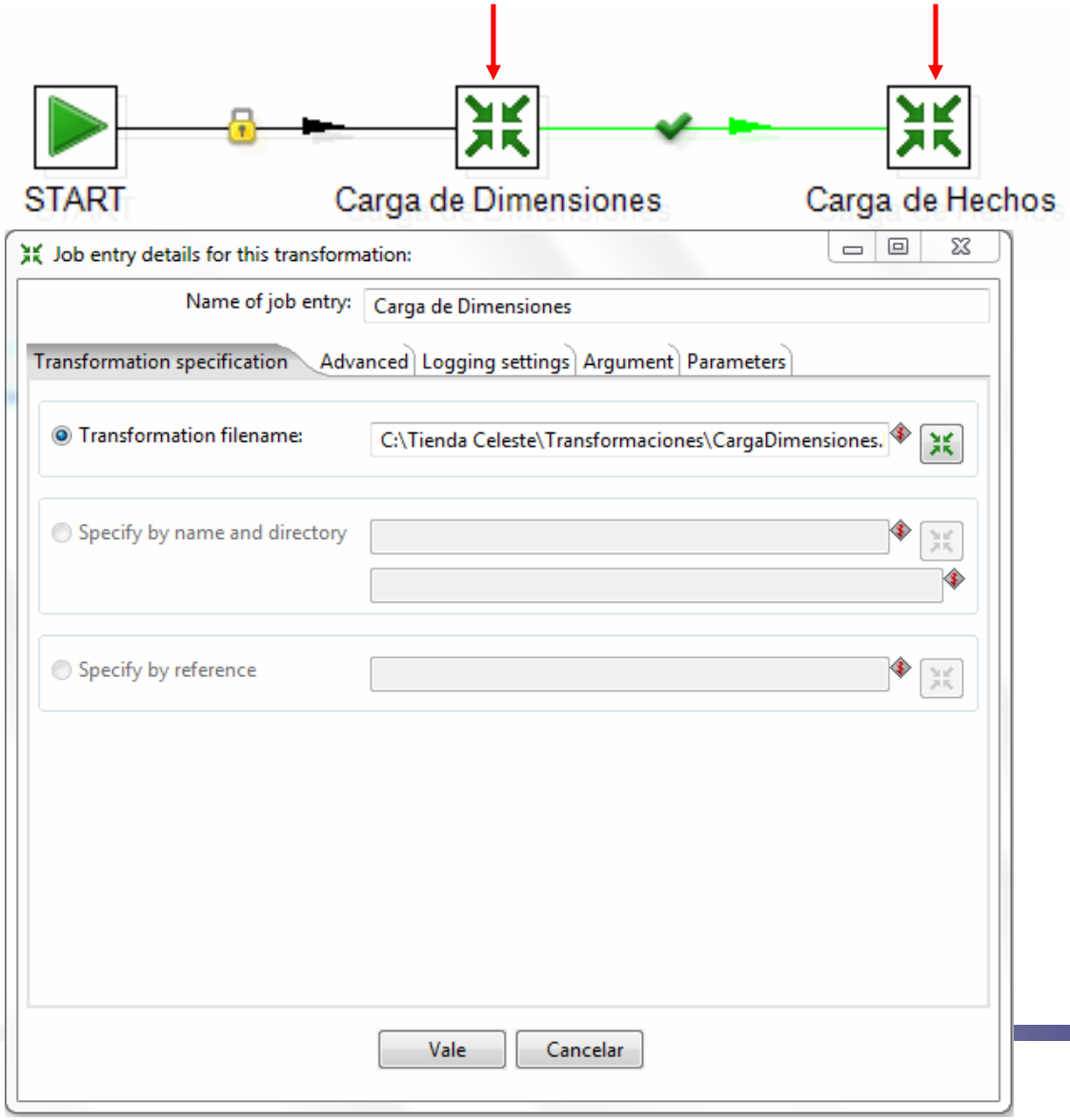
Trabajo (Job)


START
Carga de Dimensiones
Carga de Hechos

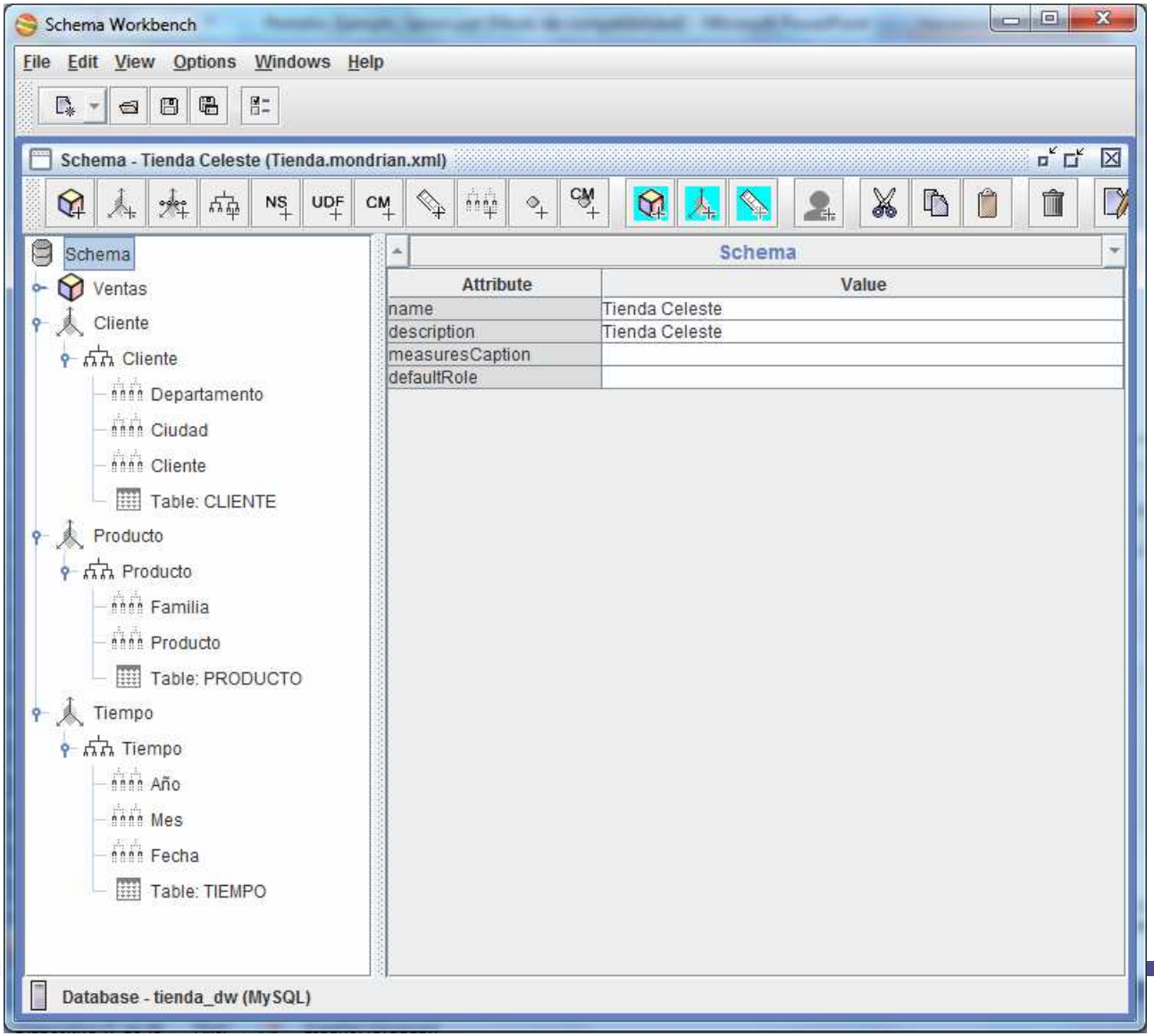

Job Scheduling
Repeat:
Type:
No Scheduling
Interval in seconds:
0
Interval in minutes:
60
Time of day:
12
0
Day of week:
Monday
Day of month:
1
Vale
Cancelar


Job Scheduling
Repeat:
Type:
No Scheduling
Interval in seconds:
Interval in minutes:
Time of day:
Day of week:
Monday
Day of month:
1
No Scheduling
Interval
Daily
Weekly
Monthly
Vale
Cancelar

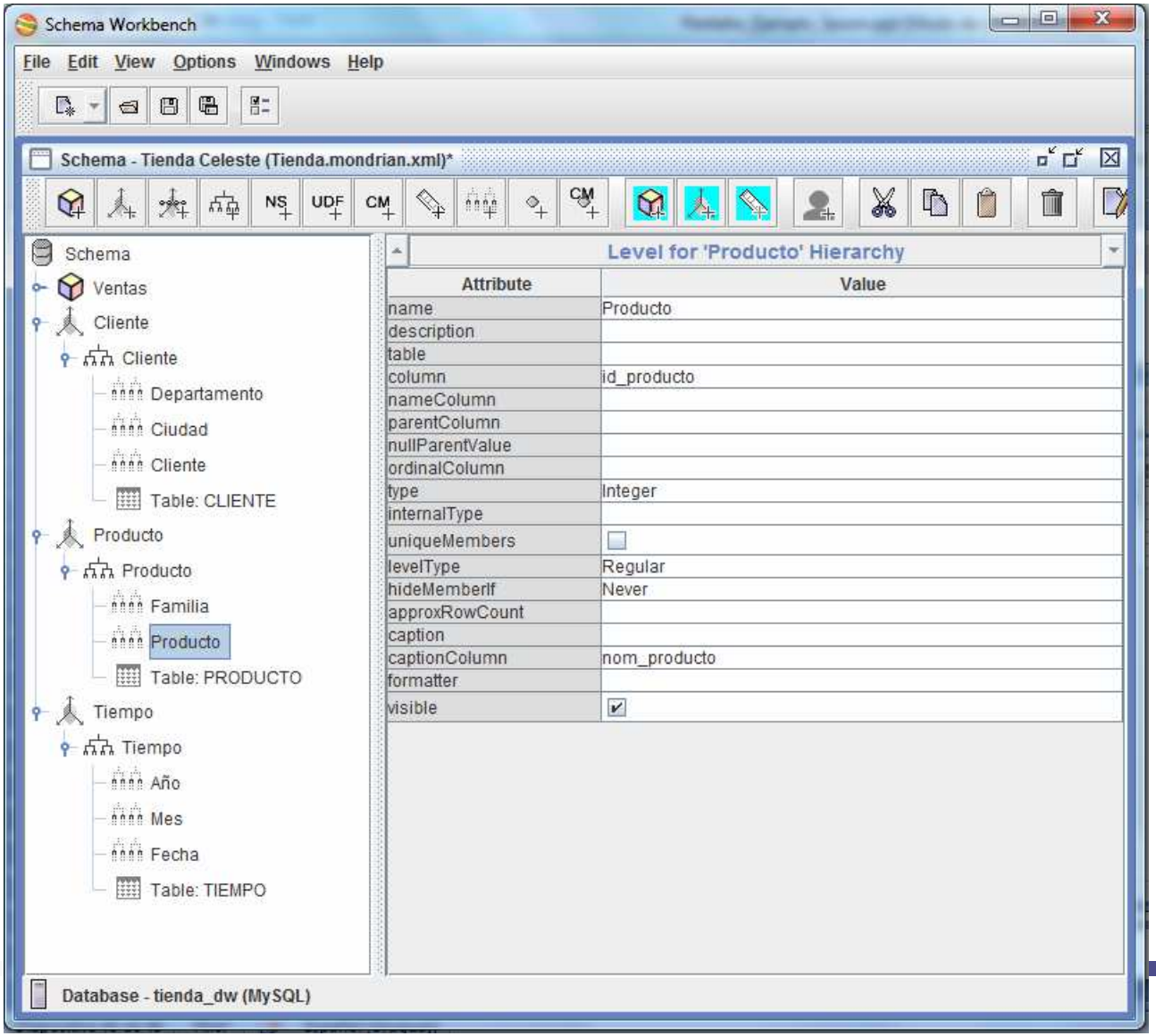
Trabajo: Step Transformación



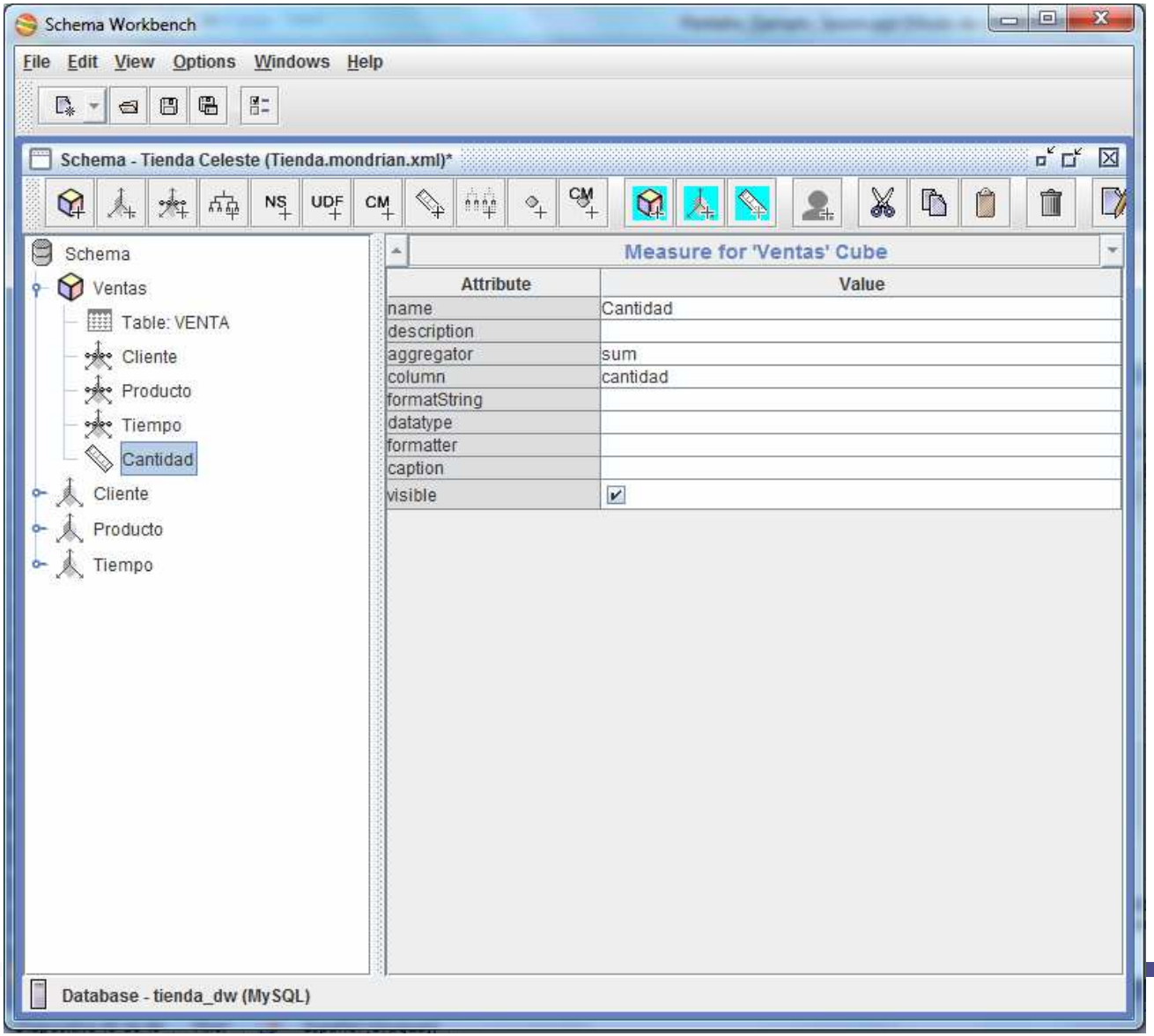
Diseño de Cubos



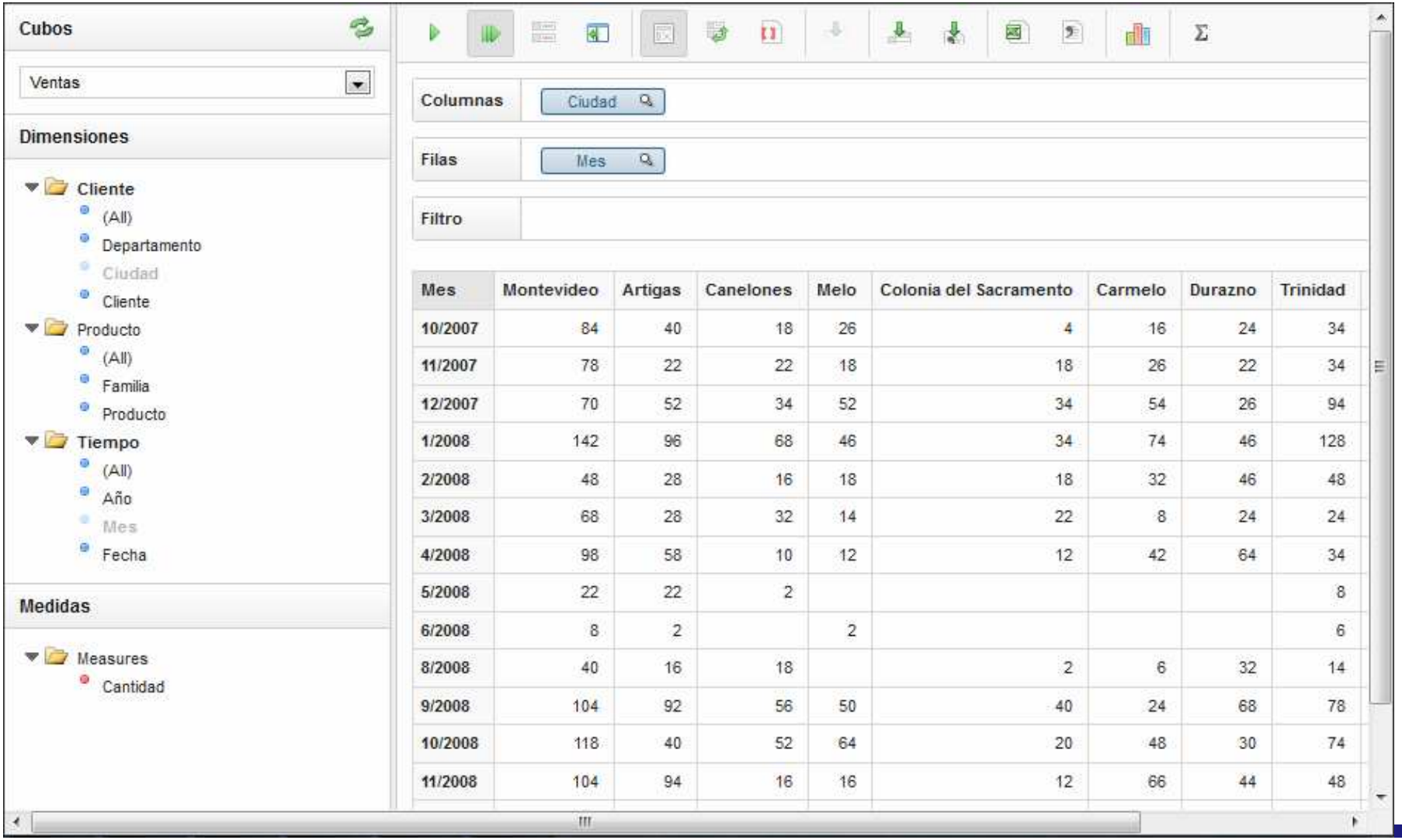
Diseño de Cubos



Diseño de Cubos

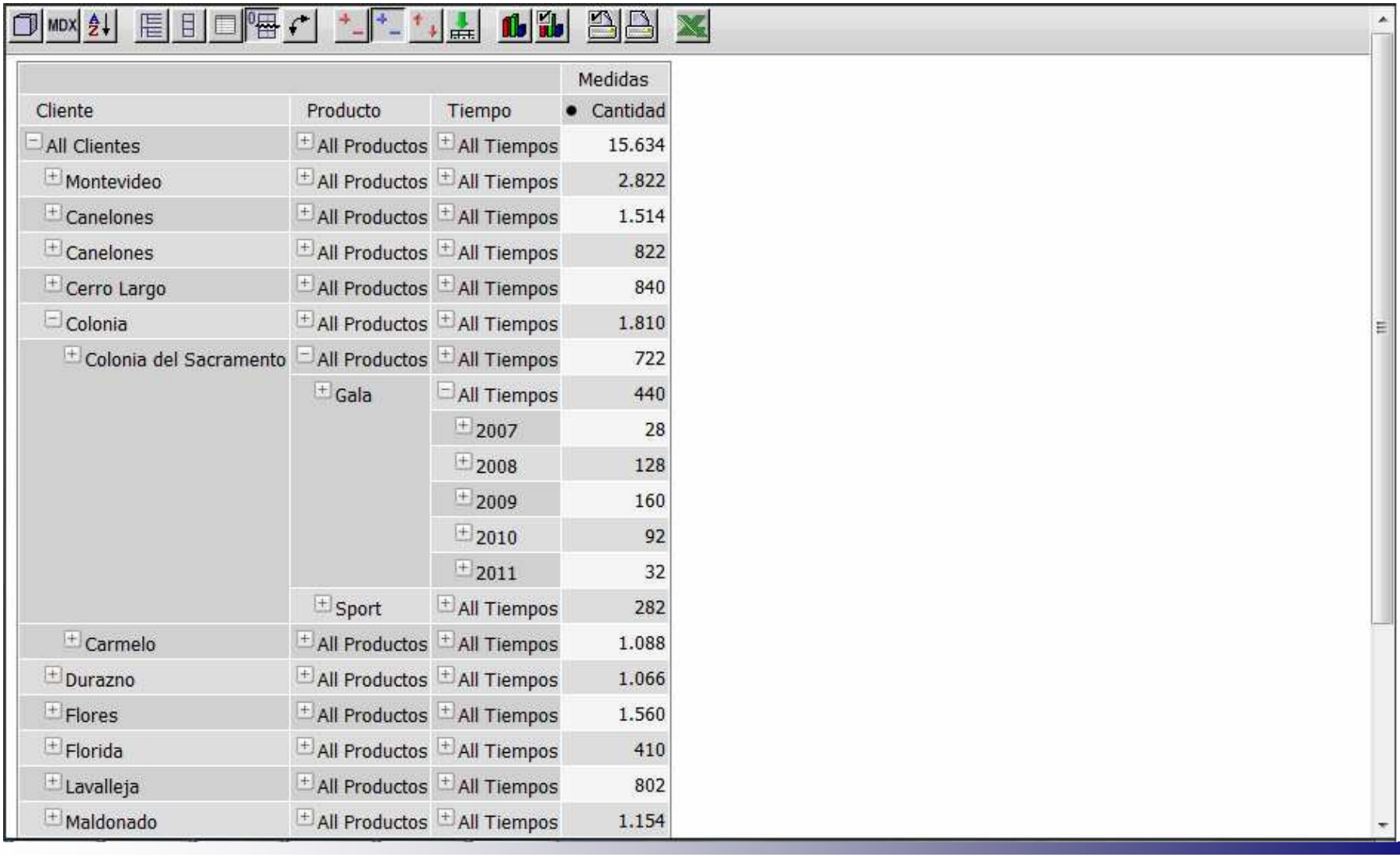


Herramienta de Análisis: Saiku



Mes	Montevideo	Artigas	Canelones	Melo	Colonia del Sacramento	Carmelo	Durazno	Trinidad
10/2007	84	40	18	26	4	16	24	34
11/2007	78	22	22	18	18	26	22	34
12/2007	70	52	34	52	34	54	26	94
1/2008	142	96	68	46	34	74	46	128
2/2008	48	28	16	18	18	32	46	48
3/2008	68	28	32	14	22	8	24	24
4/2008	98	58	10	12	12	42	64	34
5/2008	22	22	2					8
6/2008	8	2		2				6
8/2008	40	16	18		2	6	32	14
9/2008	104	92	56	50	40	24	68	78
10/2008	118	40	52	64	20	48	30	74
11/2008	104	94	16	16	12	66	44	48

Herramienta de Análisis: JPivot



			Medidas
Cliente	Producto	Tiempo	Cantidad
All Clientes	All Productos	All Tiempos	15.634
Montevideo	All Productos	All Tiempos	2.822
Canelones	All Productos	All Tiempos	1.514
Canelones	All Productos	All Tiempos	822
Cerro Largo	All Productos	All Tiempos	840
Colonia	All Productos	All Tiempos	1.810
Colonia del Sacramento	All Productos	All Tiempos	722
	Gala	All Tiempos	440
		2007	28
		2008	128
		2009	160
		2010	92
		2011	32
	Sport	All Tiempos	282
Carmelo	All Productos	All Tiempos	1.088
Durazno	All Productos	All Tiempos	1.066
Flores	All Productos	All Tiempos	1.560
Florida	All Productos	All Tiempos	410
Lavalleja	All Productos	All Tiempos	802
Maldonado	All Productos	All Tiempos	1.154