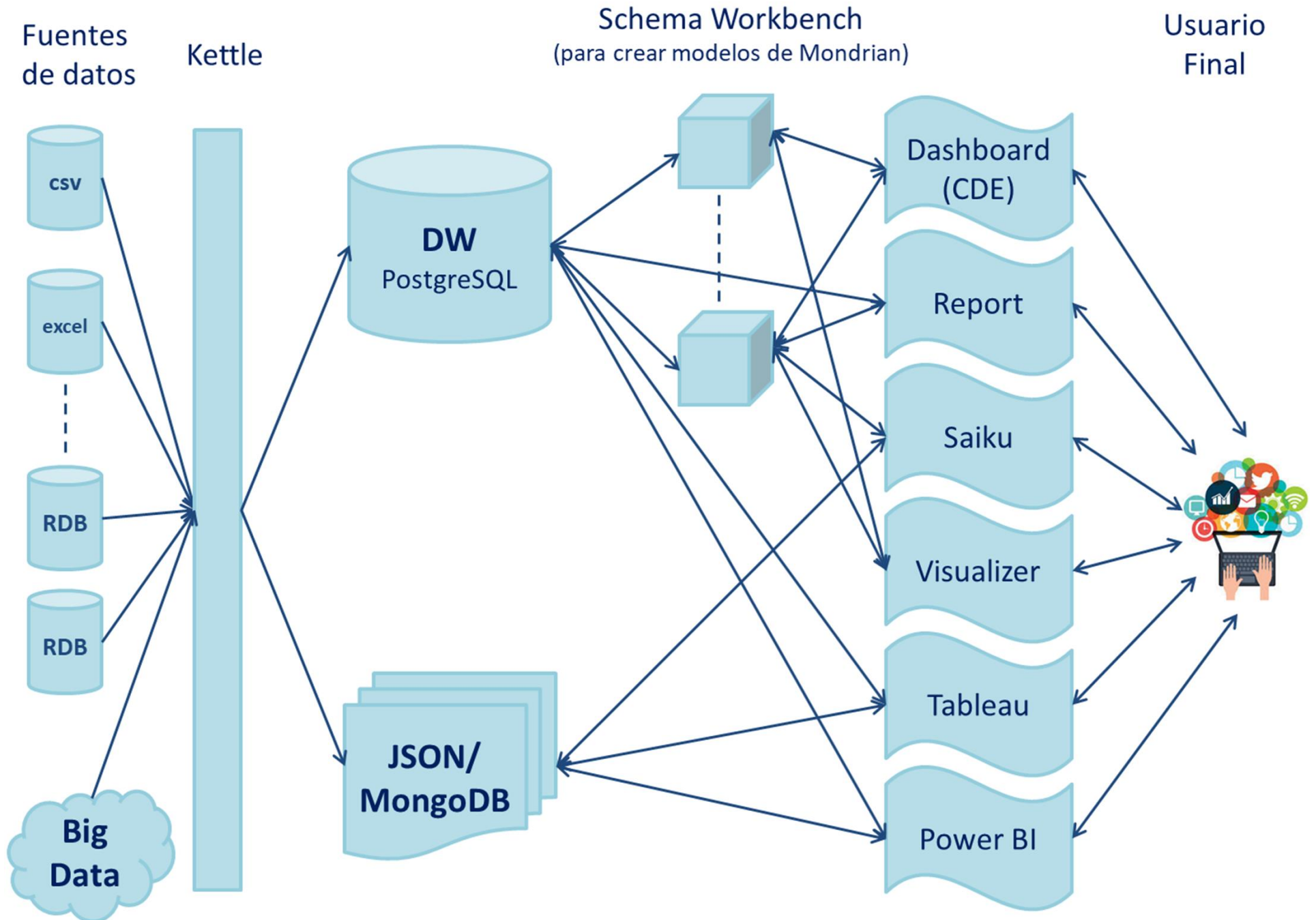




# Diseño y Construcción de Data Warehouse Caso de Estudio

*Instituto de Computación - Facultad de Ingeniería  
Edición 2021*

# Integración de las herramientas



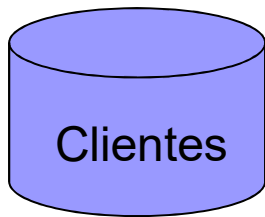


# Planteo de la realidad

- 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

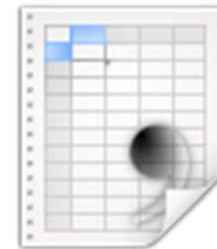
- La empresa tiene diferentes formas de almacenamiento de datos.
  - Clientes: Base de Datos
  - Productos: Planilla Excel (.xls)
  - Ventas: Planilla (.csv)



Base de Datos



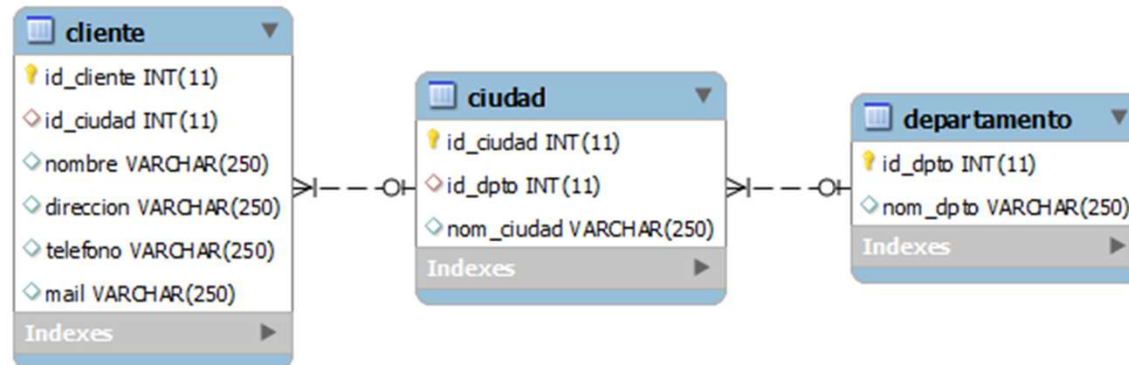
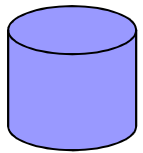
Productos (.xls)



Ventas (.csv)

# Fuentes

## Cientes



## 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
	...	...	...	...

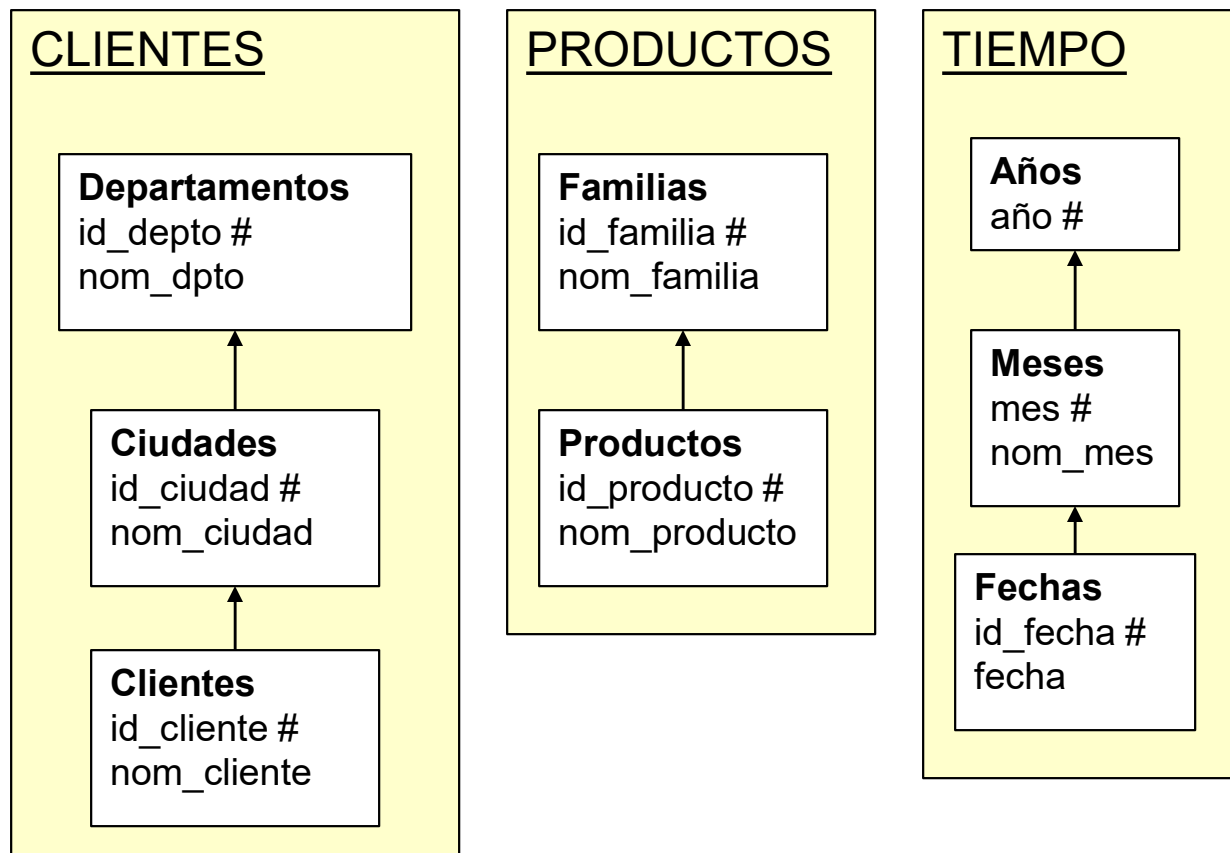
## Ventas



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

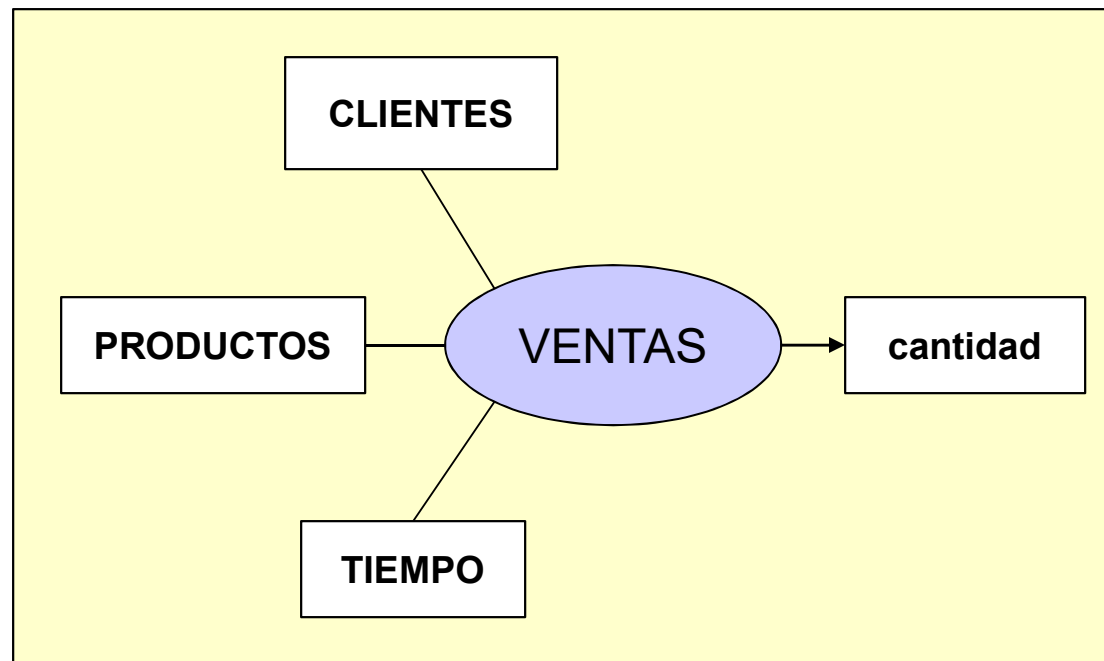
# Diseño Conceptual

## Dimensiones y Jerarquías



# Diseño Conceptual

## Relación Dimensional



# Diseño Conceptual

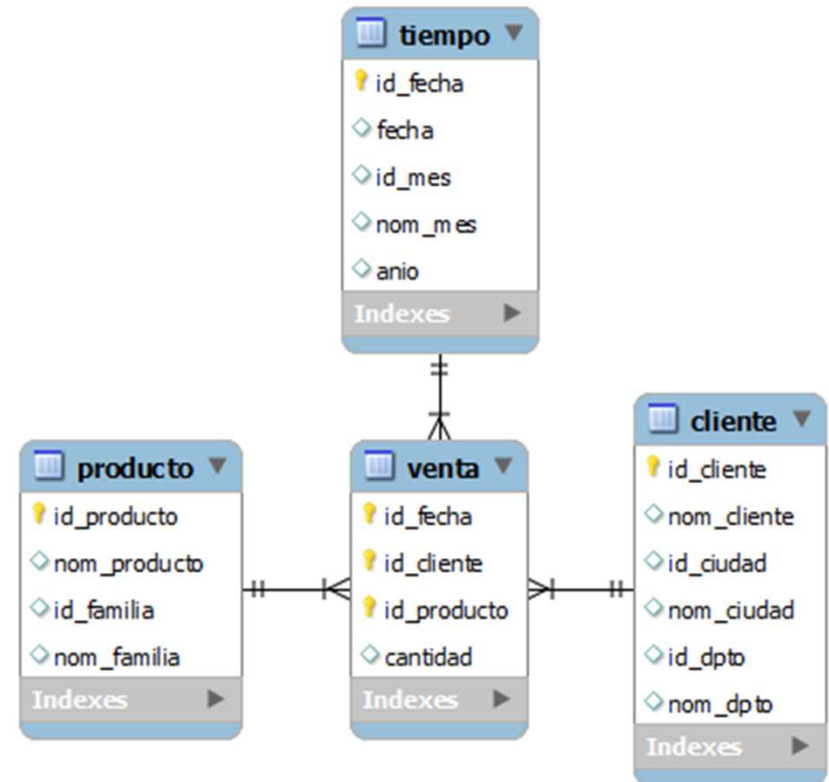
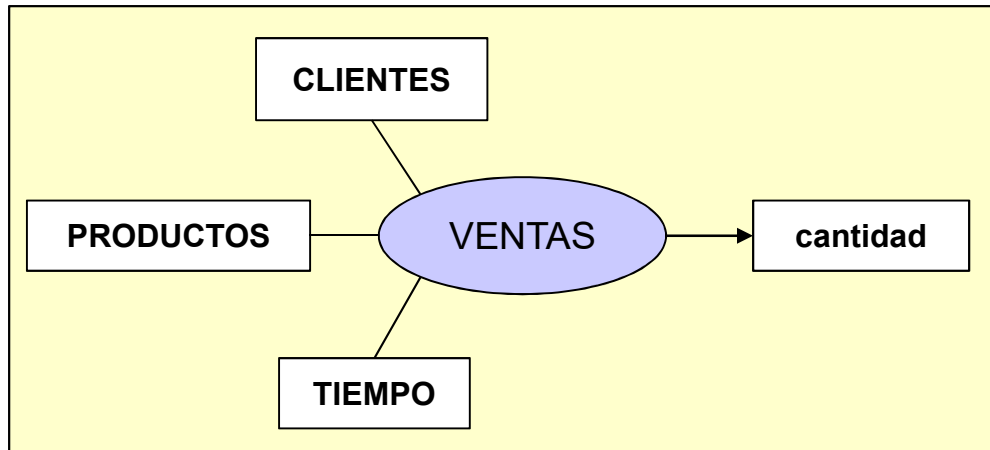
## Estudio de Aditividad

<b>Dimensión \ Medida</b>		<b>Cantidad</b>
Tiempo	<b>Fechas → Meses</b>	+
	<b>Meses → Años</b>	+
	<b>Años → ALL</b>	+
Productos	<b>Productos → Familias</b>	+
	<b>Familias → ALL</b>	+
Clientes	<b>Clientes → Ciudades</b>	+
	<b>Ciudades → Departamentos</b>	+
	<b>Departamentos → ALL</b>	+



# 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

The screenshot displays the 'Insert / Update' step configuration in Pentaho Data Integration. The step is named 'Insertar en Tabla' and is connected to the 'tienda\_dw' database. The target table is 'tiempo'. The 'Edit...' button for the connection is circled in red. Below the configuration, there are two tables: 'The key(s) to look up the value(s):' and 'Update fields:'. The 'Update fields' table shows that the 'id\_fecha' field is not updated (N), while 'fecha', 'id\_mes', 'nom\_mes', and 'anio' are updated (Y).

Steps

- Input
- Output
  - Automatic Documentator
  - Delete
  - Insert / Update**
  - JSON Output
  - LDAP Output
  - Microsoft Access Output
  - Microsoft Excel Output
  - Microsoft Excel Writer
  - Pentaho Reporting Output
  - Properties Output
  - RSS Output
  - S3 File Output
  - SQL File Output
  - Salesforce Delete
  - Salesforce Insert
  - Salesforce Update
  - Salesforce Upsert
  - Serialize to file
  - Synchronize after merge
  - Table output

Insert / Update

Step name: Insertar en Tabla

Connection: tienda\_dw (Edit... New... Wizard...)

Target schema: [Browse...]

Target table: tiempo (Browse...)

Commit size: 100

Don't perform any updates:

The key(s) to look up the value(s):

#	Table field	Comparator	Stream field1	Stream field2
1	id_fecha	=	id_fecha	

Update fields:

#	Table field	Stream field	Update
1	id_fecha	id_fecha	N
2	fecha	fecha	Y
3	id_mes	id_mes	Y
4	nom_mes	mes	Y
5	anio	year	Y

ampos

Insertar en Tabla Tiempo

Help OK Cancel SQL

# Carga de Dimensión: Tiempo

Database Connection

General  
Advanced  
Options  
Pooling  
Clustering

Connection Name:  
tienda\_dw

Connection Type:  
Oracle RDB  
Palo MOLAP Server  
Pentaho Data Services  
PostgreSQL  
Redshift  
Remedy Action Request System  
SAP ERP System  
SQLite  
SparkSQL  
Sybase  
SybaseIQ  
Teradata  
UniVerse database  
Vertica  
Vertica 5+  
dBase III, IV or 5

Settings  
Host Name:  
localhost  
Database Name:  
tienda\_DW  
Port Number:  
5433  
User Name:  
postgres  
Password:  
.....

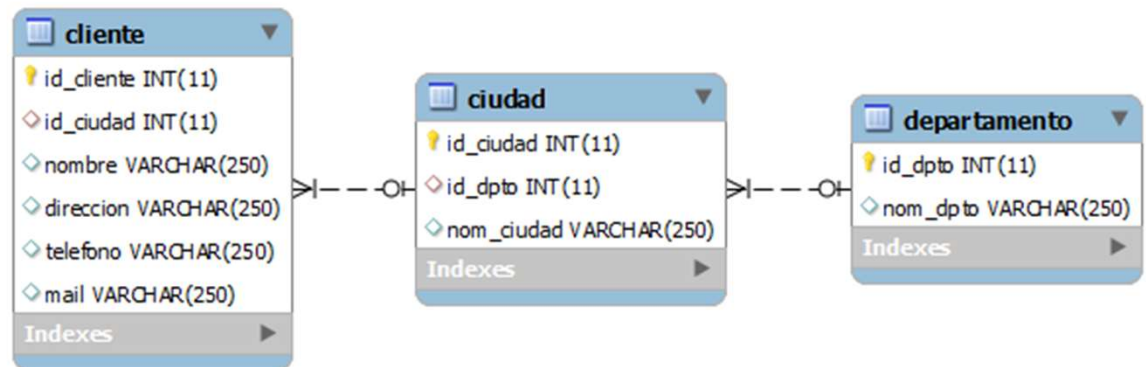
Access:  
Native (JDBC)  
ODBC  
JNDI

Test Feature List Explore

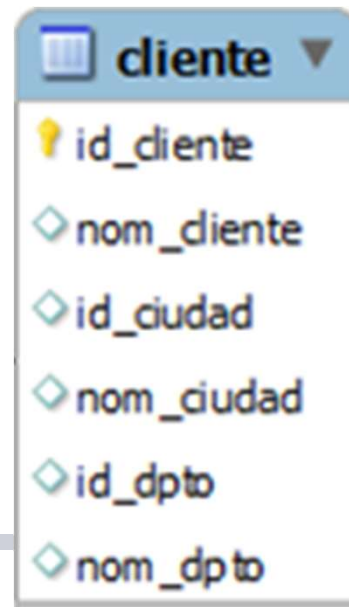
OK Cancel

# Carga de Dimensión: Cliente

□ BD Fuente (BD):



□ Tabla en el DW:



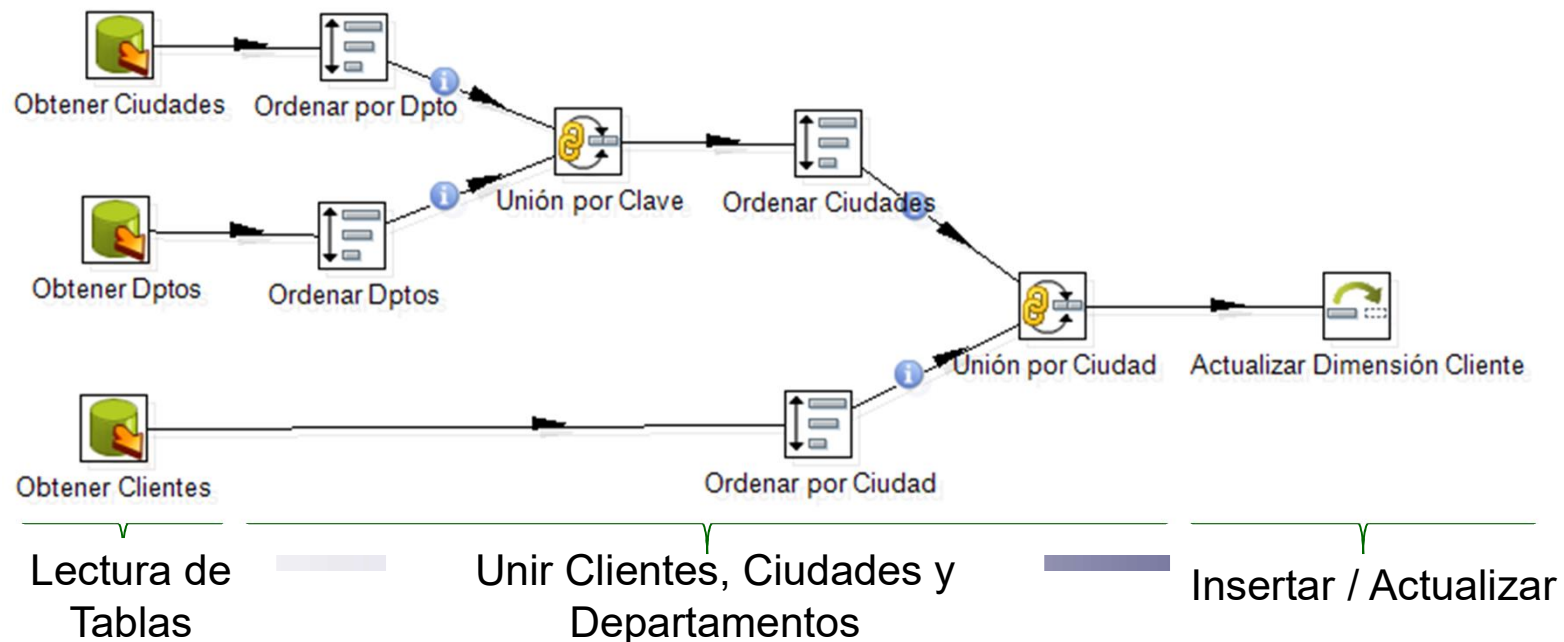


# Carga de Dimensión: Cliente

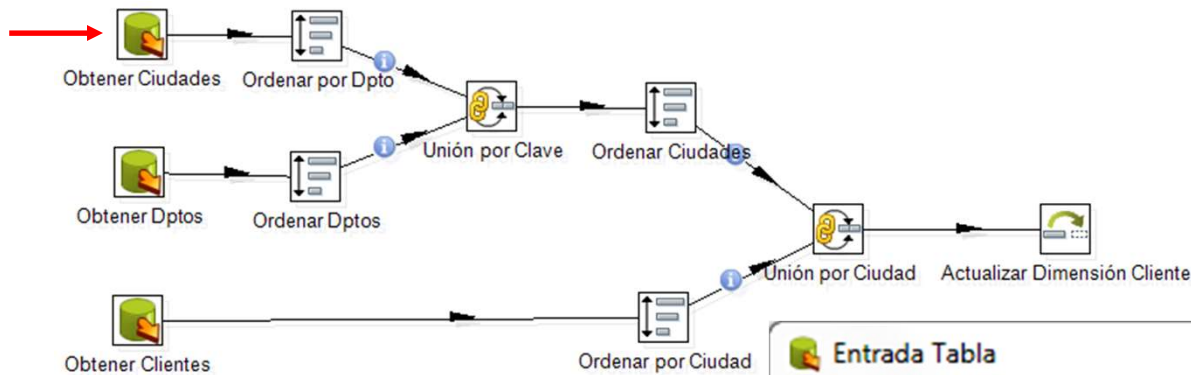
## ■ Dimensión *Cliente*:

- 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»

**Entrada Tabla**

Nombre paso:

Conexión:

SQL

```
SELECT id_ciudad, id_dpto, nom_ciudad FROM ciudad
```

Line 1 Column 0

Enable lazy conversion

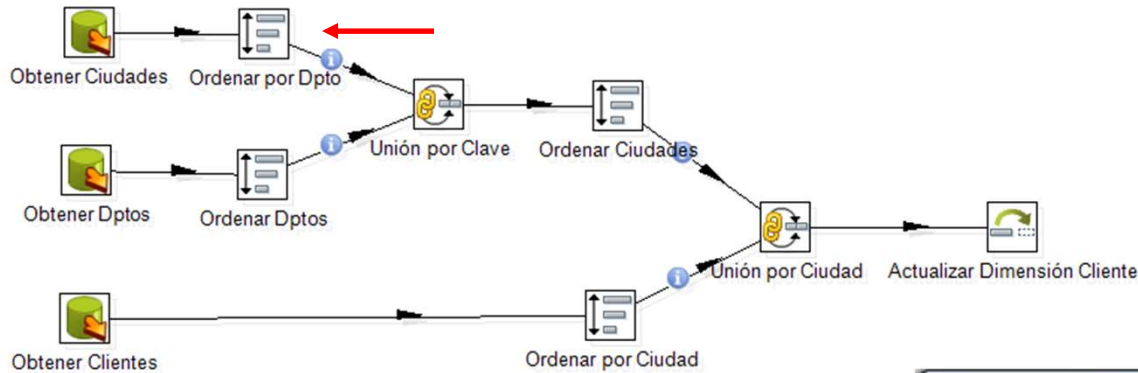
¿Reemplazar variables en

Insertar datos del paso

¿Ejecutar para cada fila?

Limitar tamaño

# Carga de Dimensión: Cliente



Cliente: *Step*  
«*Ordenar por Dpto.*»

The screenshot shows the 'Ordenar filas' dialog box with the following configuration:

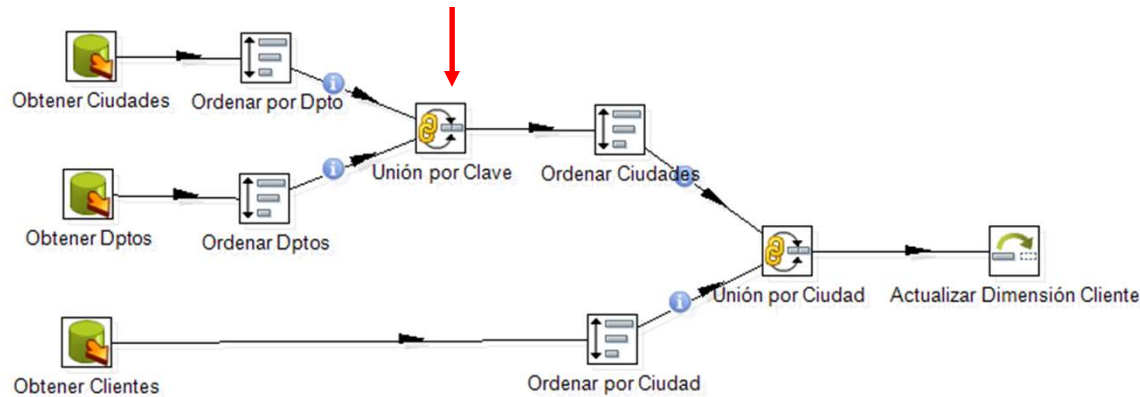
- Nombre paso: Ordenar por Dpto
- Directorio ordenación: %%java.io.tmpdir%%
- Prefijo para ficheros: out
- Tamaño de ordenación: 1000000
- Free memory threshold: (empty)
- ¿Comprimir ficheros:
- Only pass unique rows?:

Campos:

#	Nombre Campo	Ascendente	Case sensitive compare?
1	id_dpto	S	N

Buttons: Vale, Cancelar, Traer Campos

# Carga de Dimensión: Cliente



Cliente: *Step*  
«Unión por Clave»

**Unión por clave**

Nombre de paso: Unión por Clave

Primer Paso: Ordenar por Dpto

Segundo Paso: Ordenar Dptos

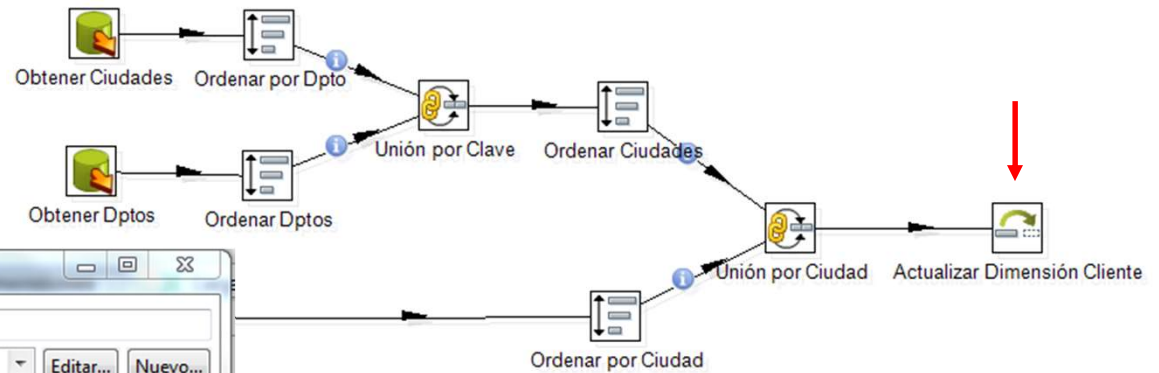
Tipo Unión: INNER

Claves de primer paso:			Claves de segundo paso:		
#	Campo clave		#	Campo clave	
1	id_dpto		1	id_dpto	

Obtener campos clave      Obtener campos clave

Vale      Cancelar

# Carga de Dimensión: Cliente



Insertar/Actualizar

Nombre de paso: Actualizar Dimensión Cliente

Conexión: tienda\_dw

Esquema destino:

Tabla destino: cliente

Tamaño transacción (commit): 100

No realizar actualizaciones:

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

#	Campo de tabla	Comparador	Campo1	Campo2
1	id_cliente	=	id_cliente	

Campos de actualización:

#	Campo de tabla	Campo de Flujo	Actualizar
1	id_cliente	id_cliente	N
2	nom_cliente	nombre	Y
3	id_ciudad	id_ciudad	Y
4	nom_ciudad	nom_ciudad	Y
5	id_dpto	id_dpto	Y
6	nom_dpto	nom_dpto	Y

Vale Cancelar SQL

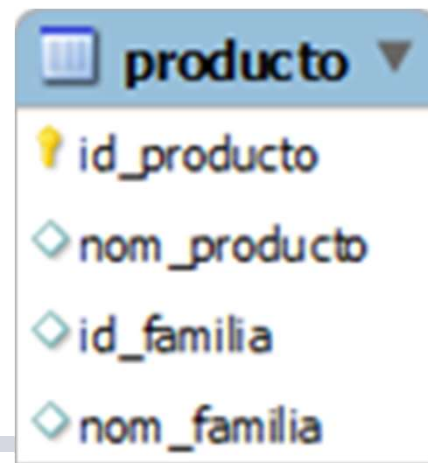
Cliente: Step  
«Actualizar dimensión  
Cliente»

# 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:

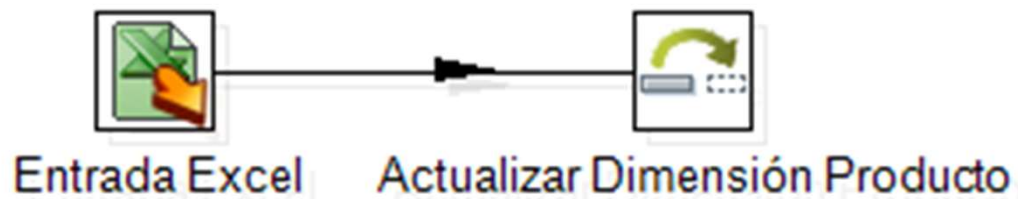


A screenshot of a data warehouse table definition for 'producto'. The table is shown in a dropdown menu format. The table name 'producto' is at the top with a dropdown arrow. Below it, four fields are listed: 'id\_producto' (marked with a yellow key icon), 'nom\_producto' (marked with a blue diamond icon), 'id\_familia' (marked with a blue diamond icon), and 'nom\_familia' (marked with a blue diamond icon).

producto
id_producto
nom_producto
id_familia
nom_familia

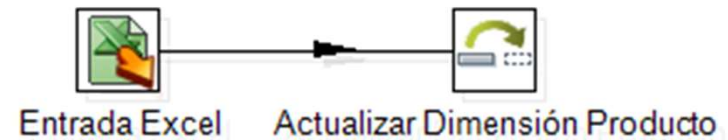
# Carga de Dimensión: Producto

- 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»



The screenshot shows the 'Entrada Excel' configuration window. The title bar reads 'Entrada Excel'. The 'Nombre paso' field contains 'Entrada Excel'. The window has several tabs: 'Ficheros', 'Hojas', 'Contenido', 'Manejador de Error', 'Campos', and 'Additional output fields'. The 'Ficheros' tab is active. It contains the following fields and controls:

- 'Fichero o directorio': A text input field with 'Añadir' and 'Examinar...' buttons to its right.
- 'Expresión Regular': A text input field.
- 'Exclude Regular Expression': A text input field with a red 'X' icon to its right.
- 'Ficheros seleccionados:': A table with the following content:

#	Fichero/Directorio
1	C:\Tienda Celeste\Fuentes\productos.xls
2	

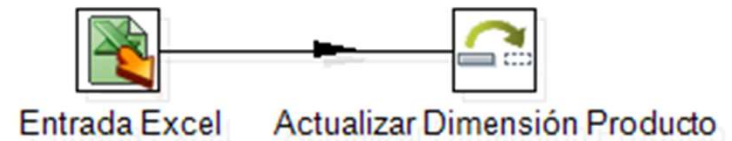
Buttons 'Eliminar' and 'Editar' are to the right of the table.
- 'Aceptar nombres de fichero de pasos anteriores': A section with a checkbox 'Aceptar nombres de fichero de pasos' (unchecked). Below it are two dropdown menus: 'Paso desde el que se obtienen los' and 'Campo de entrada a utilizar como'.
- 'Mostrar Fichero(s)...': A button.

At the bottom of the window are three buttons: 'Vale', 'Previsualizar filas', and 'Cancelar'.



# Carga de Dimensión: Producto

Producto: *Step* «Entrada Excel»



Entrada Excel

Nombre paso Entrada Excel

Ficheros Hojas Contenido Manejador de Error Campos Additional output fields

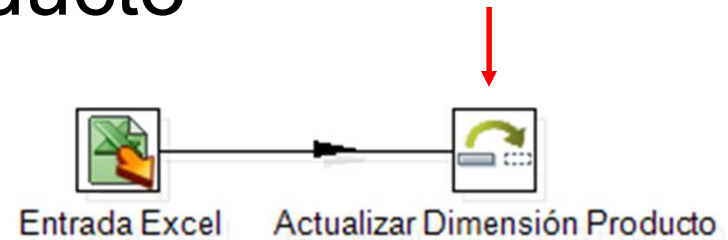
#	Nombre	Tipo	Longitud	Precisión	Tipo de poda	Repetir	Formato	Moneda	Decimal	Agrupamiento
1	Código del Producto	Number	-1	-1	ninguno	N				
2	Nombre del Producto	String	-1	-1	ninguno	N				
3	Código de la Familia	String	-1	-1	ninguno	N				
4	Nombre de la Familia	String	-1	-1	ninguno	N				

Obtener campos de cabecera...

Vale Previsualizar filas Cancelar

# Carga de Dimensión: Producto

Producto: *Step «Actualizar Dimensión Producto»*



Insertar/Actualizar

Nombre de paso: Actualizar Dimensión Producto

Conexión: tienda\_dw [Editar...] [Nuevo...]

Esquema destino: [Examinar...]

Tabla destino: producto [Examinar...]

Tamaño transacción (commit): 100

No realizar actualizaciones:

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

#	Campo de tabla	Comparador	Campo1	Campo2	Obtener Campos
1	id_producto	=	Código del Producto		

Campos de actualización:

#	Campo de tabla	Campo de Flujo	Actualizar	Obtener campos de actualización
1	id_producto	Código del Producto	N	
2	nom_producto	Código del Producto	Y	
3	id_familia	Código de la Familia	Y	
4	nom_familia	Nombre de la Familia	Y	

[Edit mapping]

[Vale] [Cancelar] [SQL]

# 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:



venta	
id_fecha	
id_cliente	
id_producto	
cantidad	



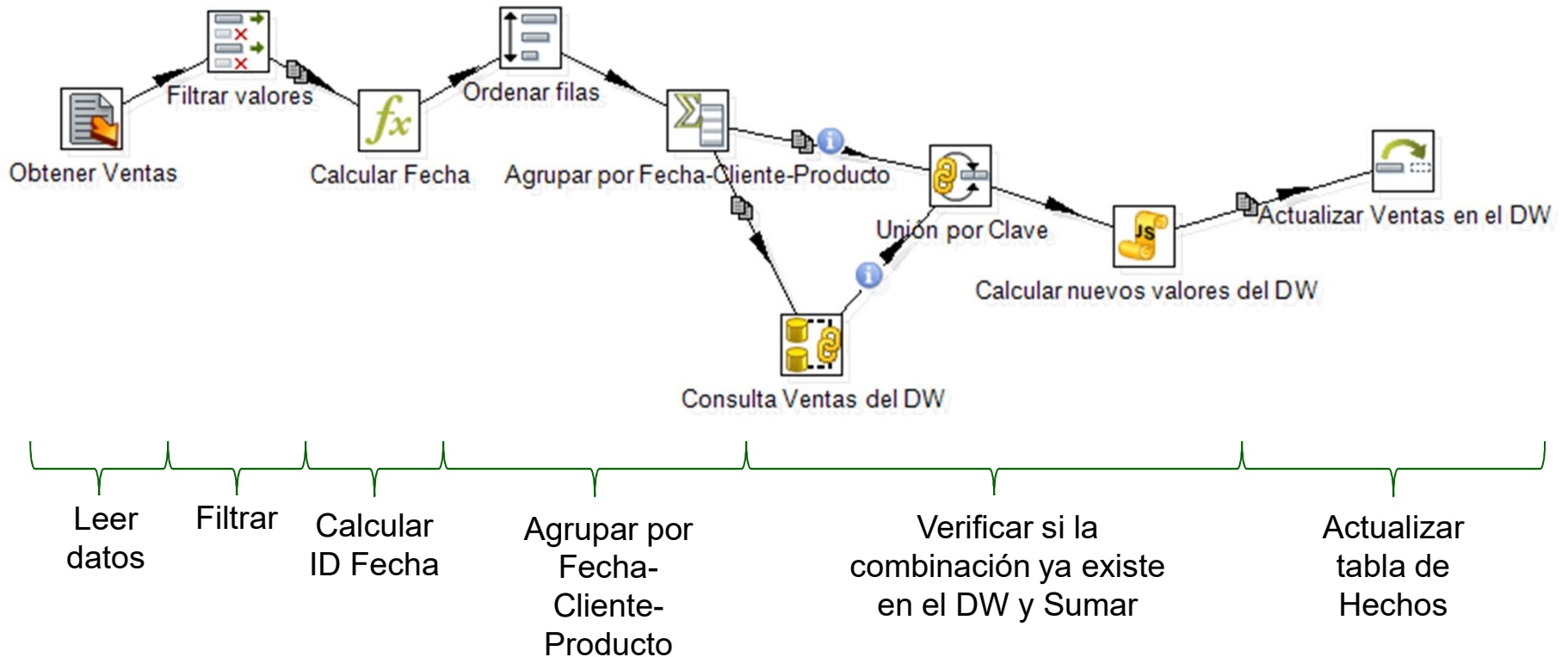
# Carga de Hechos

- Tabla de Hechos *Ventas*

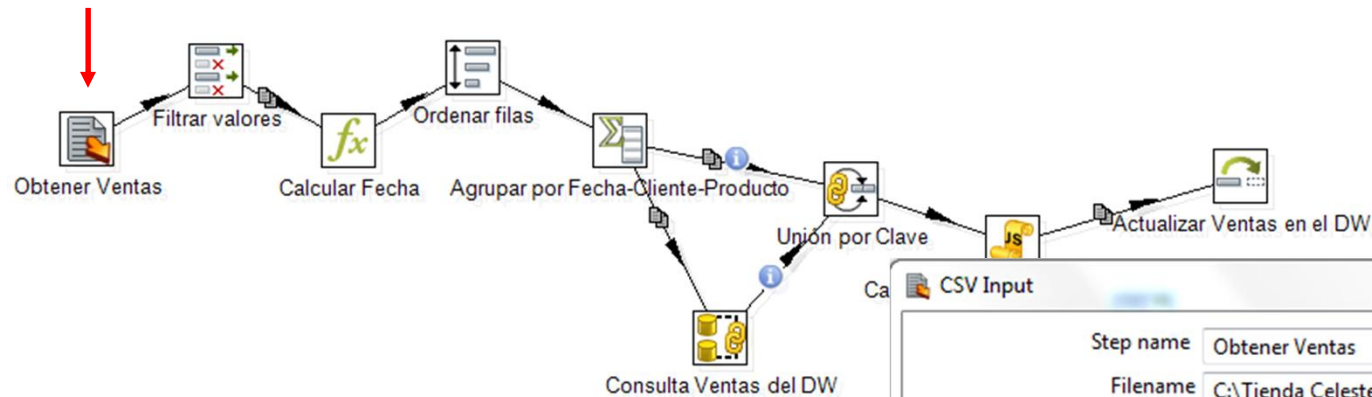
- Leer los registros de la planilla de Ventas.
- Filtrar los campos que realmente son necesarios.
- Calcular *id\_fecha* (clave foránea a dimensión Tiempo).
- Agrupar por Fecha-Cliente-Producto.
- Para las combinaciones Fecha-Cliente-Producto que ya existen en el DW, sumar la cantidad.
- Insertar en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

# Carga de Hechos: Ventas

## ■ Transformación



# Carga de Hechos



Ventas: *Step*  
«*Obtener Ventas*»

CSV Input

Step name:

Filename:

Delimiter:

Enclosure:

NIO buffer size:

Lazy conversion?

Header row present?

Add filename to result

The row number field name

Running in parallel?

New line possible in fields?

File encoding:

#	Name	Type	Format	Length	Precision
1	Fecha	Date	dd/MM/yyyy		
2	Nro. Factura	Integer	#	15	
3	Cliente	Integer	#	15	
4	Producto	Integer	#	15	
5	Cantidad	Integer	#	15	

# Carga de Hechos



Ventas: Step  
«Filtrar valores»

Selección/Renombra valores

Nombre paso: Filtrar valores

Selección & Modifica | Eliminar | Meta-información

Campos:

#	Nombre campo	Renombrar a	Longitud	Precisión
1	Fecha			
2	Cliente			
3	Producto			
4	Cantidad			

Obtener campos a seleccionar

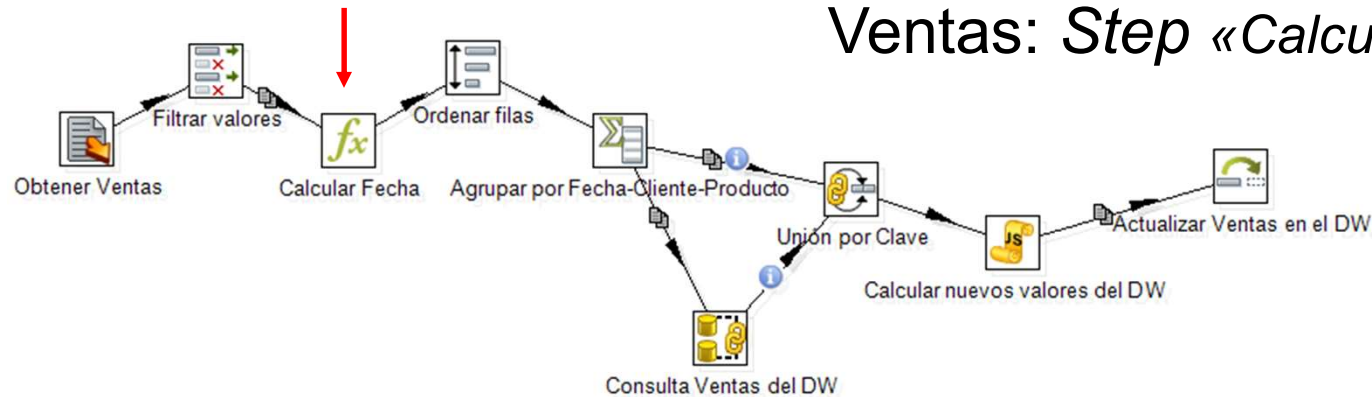
Edit Mapping

Include unspecified fields, ordered by

Vale Cancelar

# Carga de Hechos

## Ventas: Step «Calcular Fecha»



Formula

Nombre paso:

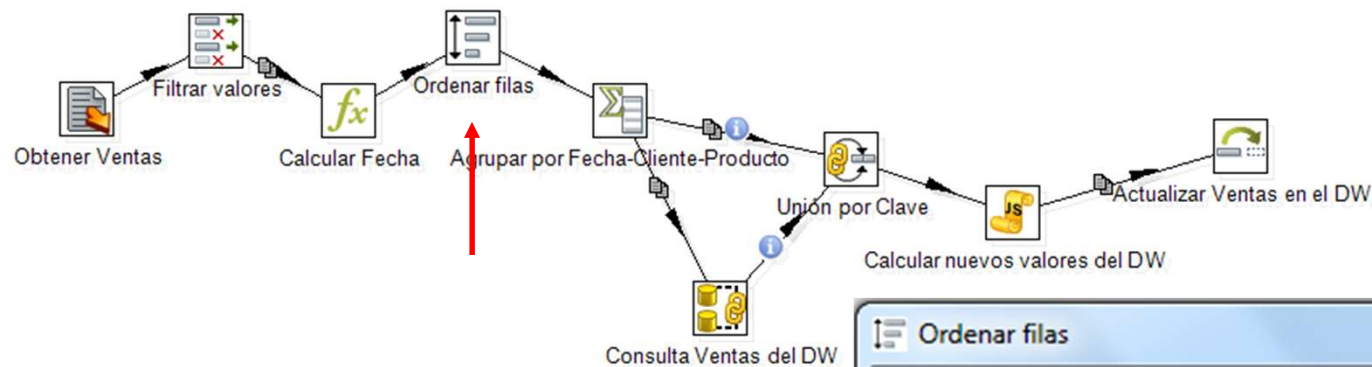
Fields:

#	New field	Formula	Value type	Le
1	id_fecha	<code>INT(Year([Fecha])*10000) + INT(Month([Fecha]) *100) + int(day([Fecha]))</code>	Number	

Vale Cancelar



# Carga de Hechos



Ventas: *Step*  
«Ordenar filas»

**Ordenar filas**

Nombre paso: Ordenar filas

Directorio ordenación: %%java.io.tmpdir%%

Prefijo para ficheros: out

Tamaño de ordenación: 1000000

Free memory threshold:

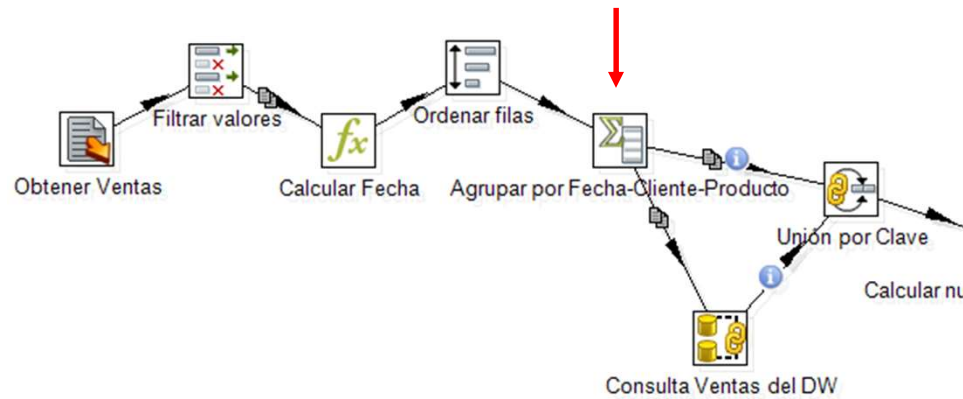
¿Comprimir ficheros:

Only pass unique rows?:

Campos:

#	Nombre Campo	Ascendente	Case sensitive compare?
1	id_fecha	S	N
2	Cliente	S	N
3	Producto	S	N

# Carga de Hechos



Ventas: Step  
«Agrupar por Fecha-Cliente-Producto»

**Agrupar**

Nombre de paso: Agrupar por Fecha-Cliente-Producto

¿Incluir todas las filas?

Directorio temporal: %%java.io.tmpdir%% Examinar...

Prefijo para ficheros temporales: grp

Añadir número de línea, reiniciar

Nombre de campo para el:

Always give back a result row

Campos que forman la agrupación:

#	Campo agrupación
1	id_fecha
2	Cliente
3	Producto

Obtener campos

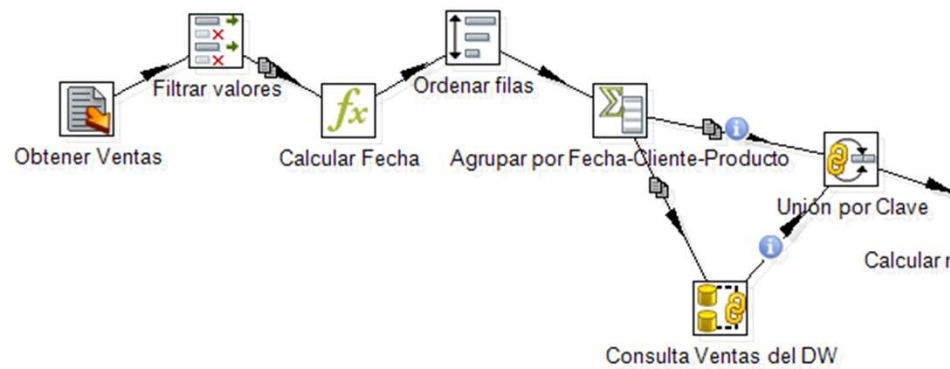
Agregados:

#	Nombre	Asunto	Tipo	Value
1	CantidadTotal	Cantidad	Suma	

Obtener campos búsqueda

Vale Cancelar

# Carga de Hechos



Ventas: *Step*  
«Consultas Ventas del DW»

Nombre de paso: Consulta Ventas del DW  
Conexión: tienda\_dw

```
SQL
SELECT cantidad as cantidad_dw
FROM venta
WHERE id_cliente = ? AND id_producto = ? AND id_fecha = ?
ORDER BY id_fecha, id_cliente, id_producto
```

Line 1 Column 0  
Número de filas a devolver: 0  
¿Outer join?   
Replace variables

Los parámetros a utilizar son:

#	Nombre campo parámetro	Tipo parámetro
1	Cliente	Integer
2	Producto	Integer
3	id_fecha	Integer

Vale Cancelar Obtener campos

# Carga de Hechos



Ventas: Step  
«Unión por Clave»

**Unión por clave**

Nombre de paso: Unión por Clave

Primer Paso: Agrupar por Fecha-Cliente-Producto

Segundo Paso: Consulta Ventas del DW

Tipo Unión: LEFT OUTER

Claves de primer paso:			Claves de segundo paso:		
#	Campo clave		#	Campo clave	
1	id_fecha		1	id_fecha	
2	Cliente		2	Cliente	
3	Producto		3	Producto	

Obtener campos clave      Obtener campos clave

Vale      Cancelar

# Carga de Hechos



Valores de Script

Nombre de paso: Calcular nuevos valores del DW

Java script functions:

- Transform Scr
- Transform Col
- Transform Fur
- Input fields
  - id\_fecha
  - Cliente
  - Producto
  - CantidadT
  - id\_fecha\_1
  - Cliente\_1
  - Producto\_1
  - CantidadT
  - cantidad\_c
- Output fields

Java script:

```
Script 1 x
//Script here
cantidad_dw_new = CantidadTotal;
if (cantidad_dw != null) {
    cantidad_dw_new += cantidad_dw;
}
```

Núm. Línea:   
 Compatibility mod  Optimization level 9

Campos

#	Nombre de campo	Renombar a	Tipo	Longitud	Precisión	Replace value 'Fieldname' or 'Rename to'
1	cantidad_dw_new		Integer			N

Vale Cancelar Obtener Variables Probar script

## 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



Insertar/Actualizar

Nombre de paso: Actualizar Ventas en el DW

Conexión: tienda\_dw

Esquema destino:

Tabla destino: venta

Tamaño transacción (commit): 100

No realizar actualizaciones:

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

#	Campo de tabla	Comparador	Campo1	Campo2
1	id_fecha	=	id_fecha	
2	id_cliente	=	Cliente	
3	id_producto	=	Producto	

Campos de actualización:

#	Campo de tabla	Campo de Flujo	Actualizar
1	id_fecha	id_fecha	N
2	id_cliente	Cliente	N
3	id_producto	Producto	N
4	cantidad	cantidad_dw_new	Y

Vale Cancelar SQL

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*

