

PEC 2 - Solución

Pregunta 1 (15%):

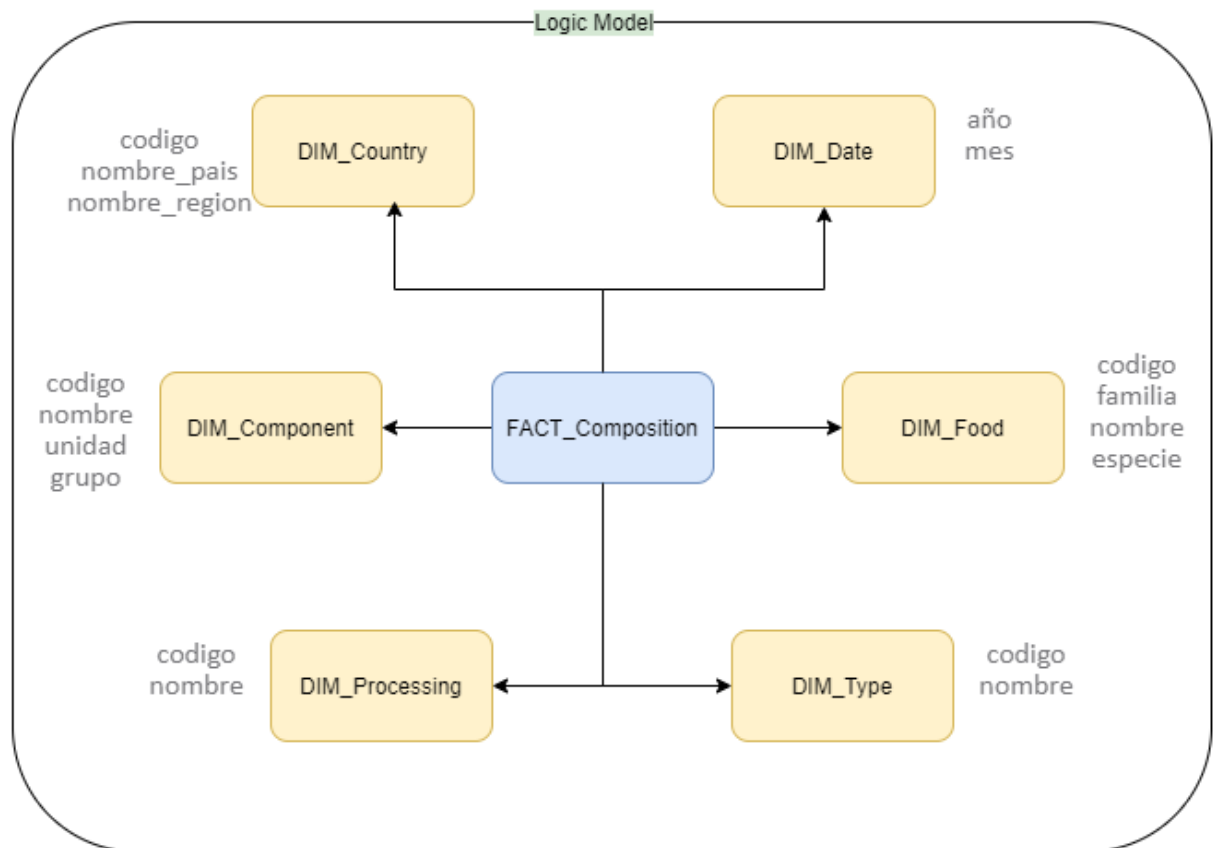
Disponemos de la tabla de hechos *FACT_Composition* que almacena información sobre las composiciones de alimentos, nutriente a nutriente, formando un esquema con forma de estrella con sus dimensiones.

Dibujad el diagrama del modelo lógico correspondiente a la tabla de hechos *FACT_COMPOSITION*, con los siguientes atributos descriptores de las dimensiones detalladas a continuación:

Dimensiones	Atributos descriptores
DIM_Country	codigo, nombre_pais, nombre_region
DIM_Date	año, mes
DIM_Food	codigo, familia, nombre, especie
DIM_Type	codigo, nombre
DIM_Processing	codigo, nombre
DIM_Component	codigo, nombre, unidad, grupo

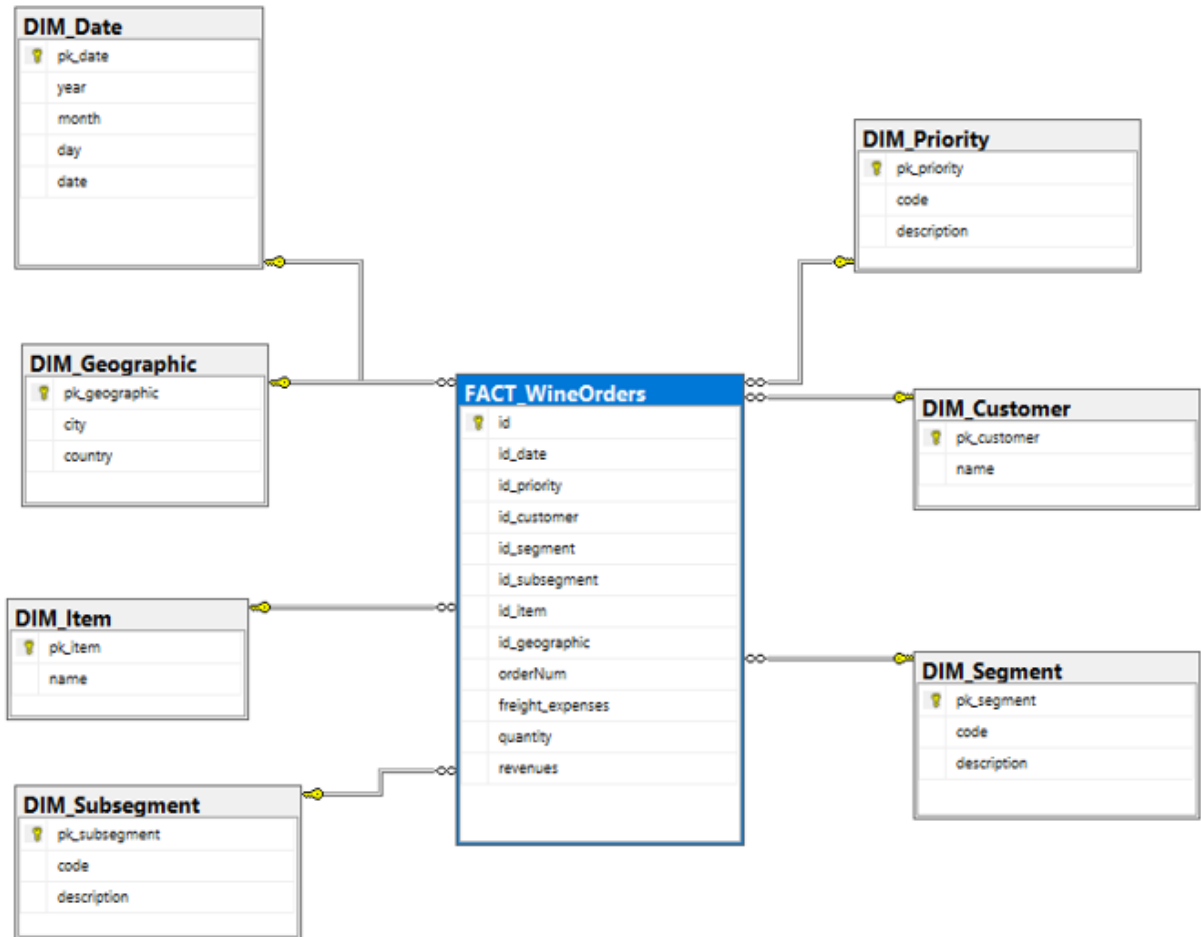
Respuesta:

El diagrama del modelo lógico correspondiente a la tabla de hechos *FACT_Compositon* con los atributos descriptores de las dimensiones es el siguiente:



Pregunta 2 (15%):

Disponemos del diseño físico siguiente:



Este modelo está basado en información sobre pedidos de venta de botellas de vino. Nos indican el significado de los siguientes campos:

- **orderNum**: Número de pedido (numérico).
- **freight_expenses**: Gastos de envío (numérico).
- **quantity**: Cantidad solicitada en el pedido (numérico).
- **revenues**: Importe de ese pedido (numérico).

A partir de esta información, **indicad si las siguientes afirmaciones son correctas o no, justificando tus respuestas.**

- El campo *[orderNum]* se podría definir como una dimensión degenerada.
- El diagrama físico no es correcto porque en la tabla de hechos faltaría añadir también como PKs los campos que se relacionan con las dimensiones:

[id_date], [id_priority], [id_customer], [id_segment], [id_subsegment], [id_item] y [id_geographic].

- c) Se dispone de 7 dimensiones (*DIM_Date*, *DIM_Priority*, *DIM_Customer*, *DIM_Segment*, *DIM_Subsegment*, *DIM_Item* y *DIM_Geographic*) y de 4 medidas (*orderNum*, *freight_expenses*, *quantity* y *revenues*).
- d) El diagrama físico es correcto porque el campo *[id]* de la tabla de hechos corresponde con una clave subrogada y éstas siempre se deben definir en las tablas de hechos.
- e) Ninguna de las anteriores es correcta.

Respuesta:

- a) **Verdadero.** Es una dimensión degenerada porque no da lugar a crear una tabla de dimensión porque no tiene atributos, pero sí que se utiliza y son útiles para identificar las instancias del hecho.
- b) **Falso.** El diagrama si es correcto porque también es habitual crear una clave subrogada como identificador único de la tabla de hechos. Este identificador se suele construir a partir de una secuencia autogenerada.
- c) **Falso.** En la respuesta a) hemos visto que *[orderNum]* puede ser una dimensión degenerada. Por tanto, tendríamos 8 dimensiones y 3 medidas.
- d) **Falso.** El diagrama físico es correcto porque el campo *[id]* de la tabla de hechos corresponde con una clave subrogada, pero no es correcto que siempre las claves subrogadas se tengan que definir en las tablas de hechos.
- e) **Falso.** La respuesta a) es correcta.

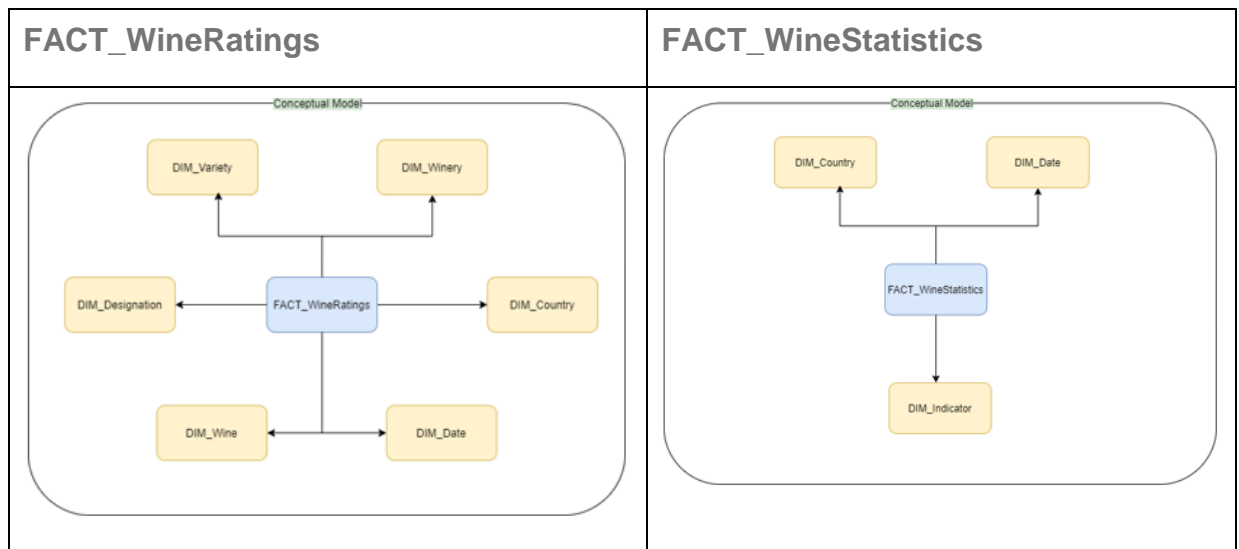
Pregunta 3 (15%):

Disponemos de dos modelos conceptuales diseñados a partir de un conjunto de fuentes con información correspondiente al sector vitivinícola.

Ambos modelos se refieren a un único *data mart*, dado que principalmente se basan en una única área temática. La información que se almacena el *data mart* está compuesta de las siguientes tablas de hechos:

- **FACT_WineRatings:** Datos de calificaciones y reseñas de vino.
- **FACT_WineStatistics:** Datos con diferentes estadísticas relacionadas con el sector vitivinícola.

A partir de estos dos modelos:



Justificad brevemente si puede existir alguna dimensión conformada en el modelo planteado.

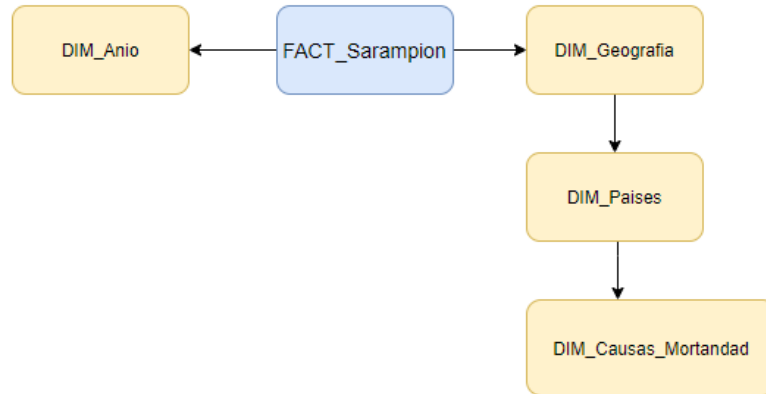
Respuesta de ejemplo:

A partir de los 2 modelos conceptuales anteriores podemos concluir que existen 2 dimensiones conformadas, la dimensión temporal (*DIM_Date*) y la dimensión país (*DIM_Country*).

Es buena práctica definir dimensiones conformadas (comunes) porque simplifica y hace más operativo el modelo final.

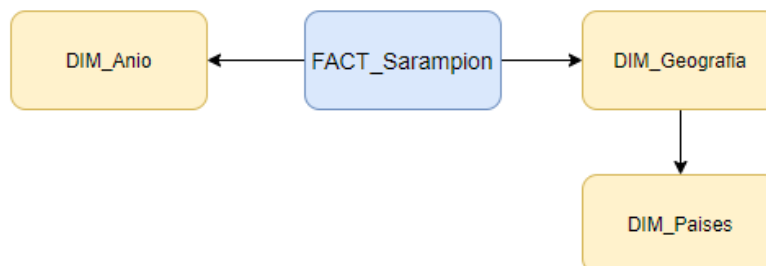
Pregunta 4 (15%):

Disponemos de la tabla de hechos FACT_SARAMPION, basada en información sobre la cobertura de inmunización. Su modelo de datos es el siguiente:



De acuerdo con este modelo, **indicad si las siguientes afirmaciones son correctas o no, justificando brevemente todas las respuestas.**

- a) Si en el diagrama del modelo conceptual se desnormaliza la información de las causas principales de mortandad en la dimensión DIM_PAISES, el diagrama del diseño conceptual de la tabla de hechos FACT_SARAMPION sería:



- a) El diseño conceptual presenta el mayor nivel de abstracción ya que es el más alejado a la representación física del modelo.
- b) La representación gráfica de su correspondiente diseño físico es un diagrama en copo de nieve.
- c) Todas las anteriores son correctas.

Respuesta:

- a) **Verdadero**, al desnormalizar se simplifica el diseño del modelo conceptual de la tabla de hechos FACT_SARAMPION, ya que la dimensión DIM_Causas_Mortandad desaparece y la información de las causas principales de mortandad se agregará en la dimensión DIM_PAISES.
- b) **Verdadero**, El nivel más bajo de abstracción describe cómo se almacenan realmente los datos y en el diseño de un almacén de datos corresponde al modelo físico, por lo que el modelo conceptual representa el nivel más alto de abstracción de los tres modelos existentes al ser el más alejado a la creación física de la tabla de hechos FACT_SARAMPION en la base de datos.
- c) **Verdadero**, según el punto 4.2 El copo de nieve del módulo 4, el diagrama de la FACT_SARAMPION es un diagrama en copo de nieve ya que las dimensiones DIM_Paises y DIM_Causas_Mortandad se representan como ramificaciones de la dimensión DIM_Geografia.
- d) **Verdadero**, todas las respuestas anteriores son correctas.

Pregunta 5 (40%):

/ La solución de esta pregunta se indica en el mismo enunciado en fuente azul **/**

A partir del fichero “ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx” hoja “Datos_tratados”, se debe diseñar, implementar y ejecutar los procesos de extracción, transformación y carga para la Transformación IN_DENUNCIAS-INFRACCIONES, siguiendo y completando las siguientes cuestiones:

- a) Completad y ejecutad el siguiente comando SQL para la creación de la tabla intermedia donde se almacenará los datos del origen “ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx” hoja “Datos_tratados”.

IN_DENUNCIAS-INFRACCIONES

```
CREATE TABLE [dbo].[STG_Denuncias_Infracciones] (  
    [provincia] [varchar](100) NULL,  
    [identificados_ertzaintza] [float] NULL,  
    [detenidos_ertzaintza] [float] NULL,  
    [denuncias_ertzaintza] [float] NULL,  
    [vehic_intercept_ertzaintza] [float] NULL,  
    [identificados_ppl] [float] NULL,  
    [detenidos_ppl] [float] NULL,  
    [denuncias_ppl] [float] NULL,  
    [vehic_intercept_ppl] [float] NULL,  
    [fecha_final] [datetime] NULL  
    ) ON [PRIMARY]  
GO
```

- a) Lectura de los ficheros.xlsx. Completad la siguiente información del paso “File Input”:

Nombre: Entrada Excel (ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx)

Componente: Microsoft Excel input

Descripción: Permite cargar datos de entrada provenientes de un fichero Excel.

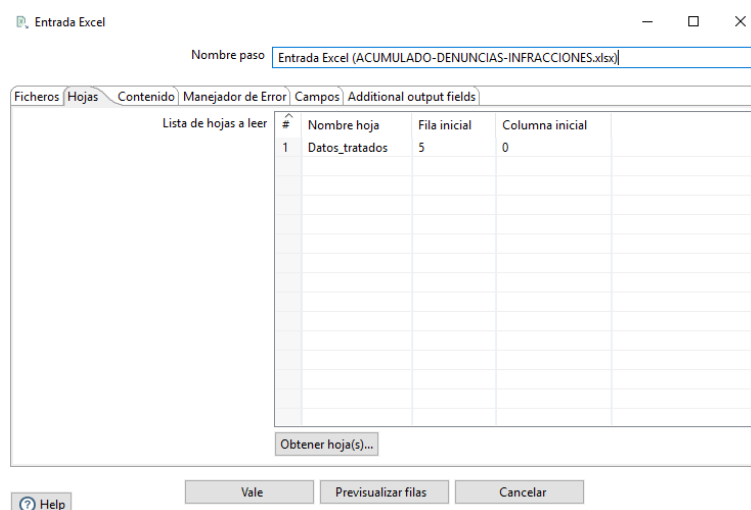
El primer paso de la transformación corresponde a la lectura del fichero origen, como se trata de un fichero XLSX se utilizará como entrada el tipo «Microsoft Excel Input». Concretamente “ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx” hoja “Datos_tratados”.

Parámetros:

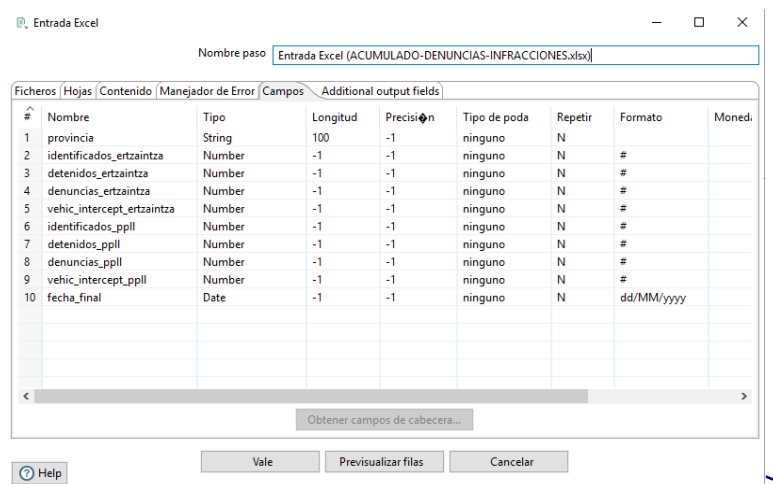
Files / File or directory: $\{\text{DIR_ENT}\}\text{ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx}$

Para facilitar la lectura del fichero se utiliza la variable de entorno «DIR_ENT»:

Sheets (Hojas) mediante “get sheet names”, hoja “Datos_tratados”, con fila inicial “start row” = 5.



Fields (campos) Mediante el botón «Get fields from header row...» se obtienen todos los campos del fichero, así como el tipo, formato y longitud del dato.



Preview: botón «*Preview rows*» (Previsualizar filas).

Rows of step: Entrada Excel (ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx) (219 rows)

#	provincia	identificados_ertaintza	detenidos_ertaintza	denuncias_ertaintza	vehic_intercept_ertaintza	identificados_ppll	detenidos_ppll	denuncias_ppll	vehic_intercept_ppll	fecha_Final
1	ARABA	10717	40	2586	15182	21599	29	2748	19250	18/06/2020
2	BIZKAIA	29955	228	6249	56139	27160	65	9209	50539	18/06/2020
3	GIPIZKOA	26051	62	4884	17246	27069	40	4473	37797	18/06/2020
4	ARABA	10708	40	2586	15180	21598	29	2748	19250	17/06/2020
5	BIZKAIA	29987	228	6249	56009	27138	65	9209	50518	17/06/2020
6	GIPIZKOA	26004	62	4882	17183	27059	40	4472	37791	17/06/2020
7	ARABA	10705	40	2585	15176	21598	29	2748	19250	16/06/2020
8	BIZKAIA	29751	228	6249	55859	27124	65	9209	50468	16/06/2020
9	GIPIZKOA	25942	62	4882	17173	27038	38	4465	37782	16/06/2020
10	ARABA	10704	40	2585	15176	21593	29	2746	19250	15/06/2020
11	BIZKAIA	29674	228	6247	55754	27106	65	9203	50436	15/06/2020
12	GIPIZKOA	25872	62	4879	17103	27022	38	4465	37770	15/06/2020
13	ARABA	10700	40	2585	15173	21573	29	2739	19250	14/06/2020
14	BIZKAIA	29587	228	6246	55674	27063	65	9202	50428	14/06/2020
15	GIPIZKOA	25775	62	4872	17003	27015	38	4465	37754	14/06/2020
16	ARABA	10652	40	2572	15134	21572	29	2738	19250	13/06/2020
17	BIZKAIA	29477	228	6236	55470	27055	65	9201	50386	13/06/2020
18	GIPIZKOA	25672	62	4862	16924	27004	38	4464	37744	13/06/2020

c) Asegurad la homogeneidad de los datos mediante la normalización de los valores de los campos tipo «*String*». Convirtiendo a mayúsculas y eliminando los espacios en blanco al inicio y al final de cada cadena.

Nombre: String Operation: Upper&Trim

Componente: String Operation

Descripción: Permite realizar las operaciones de cadena sobre un campo entrante:

- Trim (eliminar los espacios iniciales y / o finales).
- Convertir todo a mayúsculas o minúsculas, o a mayúsculas iniciales
- Agregar caracteres adicionales iniciales o finales.
- Ignorar los caracteres de escape.
- Eliminar o devolver solo dígitos numéricos.
- Eliminar los caracteres especiales.

Asegura la homogeneidad de los datos mediante la normalización de los valores de los campos tipo «String». Convirtiendo a mayúsculas y eliminando los espacios en blanco al inicio y al final de cada cadena.

Parámetros:

[illegible]

d) Ordenación ascendente de todos los campos según su colocación en la tabla Staging.

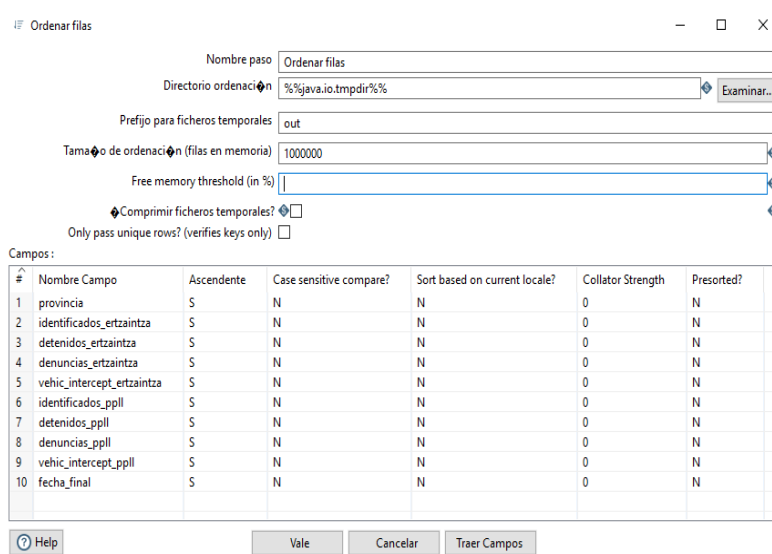
Nombre: Ordenar filas

Componente: Sort rows

Descripción: ordena las filas en función de los campos que especifique y si deben ordenarse en orden ascendente o descendente.

Ordenación ascendente de todos los campos según su colocación en la tabla Staging.

Parámetros:



#	Nombre Campo	Ascendente	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?
1	provincia	S	N	N	0	N
2	identificados_ertzaintza	S	N	N	0	N
3	detenidos_ertzaintza	S	N	N	0	N
4	denuncias_ertzaintza	S	N	N	0	N
5	vehic_intercept_ertzaintza	S	N	N	0	N
6	identificados_ppil	S	N	N	0	N
7	detenidos_ppil	S	N	N	0	N
8	denuncias_ppil	S	N	N	0	N
9	vehic_intercept_ppil	S	N	N	0	N
10	fecha_final	S	N	N	0	N

e) Cargad la información transformada en la tabla de base de datos.

Nombre: Salida Tabla (STG_Denuncias_Inflacciones)

Componente: Table Output

Descripción: carga datos en una tabla de base de datos. Es equivalente al operador de SQL INSERT.

Carga los datos resultantes de las transformaciones precedentes en la tabla intermedia del stage area: [dbo].[STG_Denuncias_Infracciones]

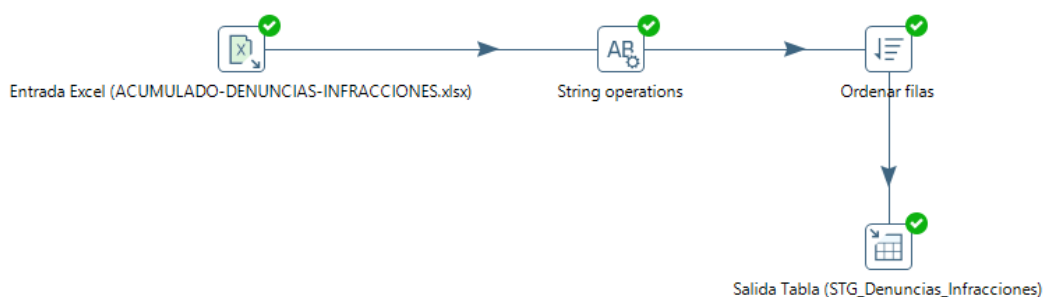
Parámetros:

Connection: Mediante la variable de entorno «CN_STAGE»

Target table: [dbo].[STG_Denuncias_Infracciones]

Truncate table : Si, marcado para posibles reprocesos.

f) Capturad de pantalla de la transformación completa, incluyendo la pestaña informativa de ejecución “*step metrics*”.



Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data											
#	Nombre paso	Numero Copia	Leído	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo
1	Entrada Excel (ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx)	0	0	219	219	0	0	0	0	Finalizado	2.4s
2	String operations	0	219	219	0	0	0	0	0	Finalizado	2.4s
3	Ordenar filas	0	219	219	0	0	0	0	0	Finalizado	2.4s
4	Salida Tabla (STG_Denuncias_Infracciones)	0	219	219	0	219	0	0	0	Finalizado	2.6s

- g) Realizad una Consulta en la Base de datos, que devuelva el número de registros de la tabla cargada. ¿Coincide con el número de registros procesados en cada paso, mostrados en “step metrics”?

Select count(*) From [dbo].[STG_Denuncias_Infracciones]

Coincide con los 219 registros del proceso completo.

- h) Realizad la consulta en la Base de Datos y capturad de resultado del Top 10 de registros sin ordenar, ¿coinciden con los 10 primeros registros ordenados ascendentemente de todos los campos según su colocación en la tabla *Staging*?

```

SELECT TOP (10) [provincia]
, [identificados_ertzaintza]
, [detenidos_ertzaintza]
, [denuncias_ertzaintza]
, [vehic_intercept_ertzaintza]
, [identificados_ppll]
, [detenidos_ppll]
, [denuncias_ppll]
, [vehic_intercept_ppll]
, [fecha_final]
FROM [dbo].[STG_Denuncias_Infracciones]
  
```

Si, Coinciden los 10 primeros registros.