Prática sesión 4 -   
DISEÑO DIMENSIONAL con pentaho

En este documento se recoge un supuesto práctico donde poder desarrollar las nociones básicas para construir un esquema multidimensional para un DWH utilizando la herramienta Pentaho Data Integration (Kettle).

Contenido

[1. Introducción 2](#_Toc71179369)

[2. Instalación Postgresql 2](#_Toc71179370)

[3. Dimensiones del modelo 6](#_Toc71179371)

[3.1 CREACIÓN DE UNA TRANSFORMACIÓN ETL PARA CARGAR Y ADMINISTRAR LOS DATOS DE LOS CENTROS COMO UNA SCD TIPO 2 7](#_Toc71179372)

[3.2 GENERACIÓN DE UNA TABLA DE DIMENSIÓN DE FECHAS 10](#_Toc71179373)

[4. Hecho del modelo 16](#_Toc71179374)

[4.1 DESARROLLO DE LA TRANSFORMACIÓN PARA CARGAR LA TABLA DE HECHOS 16](#_Toc71179375)

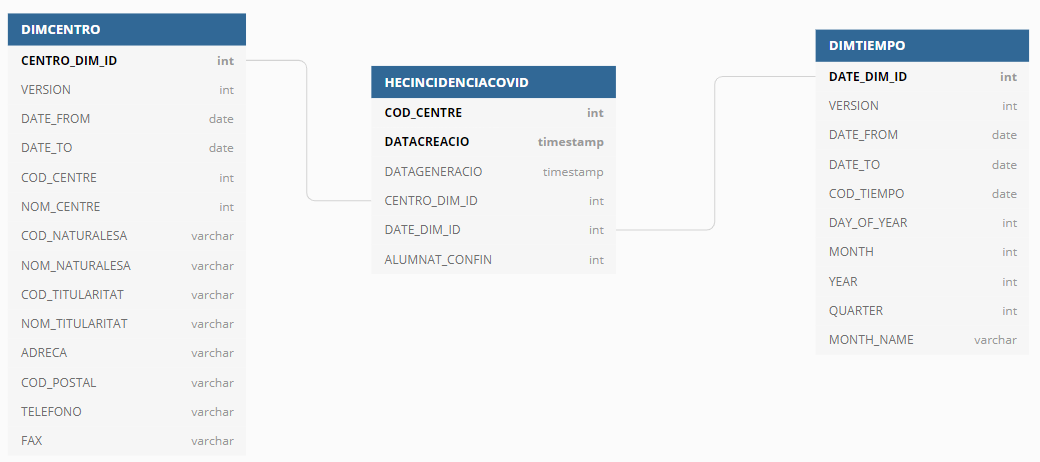
## Introducción

El objetivo de este ejercicio es crear un modelo dimensional partiendo de un dataset ofrecido por la Generalitat de Catalunya donde se recogen los datos de los casos de covid-19 que se han producido en los centros educativos a través de la aplicación Traçacovid que da soporte a los centros en el marco de protocolo de casos covid-19.

El dataset para los datos sobre las incidencias por covid reportadas por los centros se puede descargar aquí (.csv): [enlace](https://datos.gob.es/en/catalogo/a09002970-datos-covid-19-en-los-centros-educativos).

El dataset con la información de centros se puede descargar de aquí (.csv): [enlace](https://datos.gob.es/en/catalogo/a09002970-directorio-de-centros-docentes1)

Como ya se hizo en la práctica del tema 2 – NoSQL vamos a ver cómo modelar los datos para obtener, en esta ocasión, un pequeño modelo dimensional con la siguiente estructura:



## Instalación Postgresql

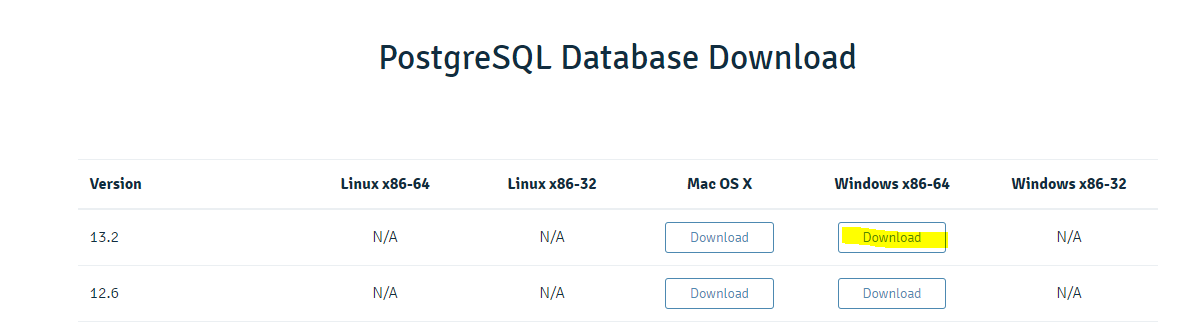
Para crear las entidades de nuestro modelo vamos a necesitar utilizar un componente del tipo Dimension lookup/update. En las propiedades de este componente tenemos que especificar la conexión a la base de datos donde queramos albergar nuestro modelo:



Si seleccionamos la opción “New” se nos mostrará una ventana donde aparecen todos los tipos de conectores que dispone pentaho

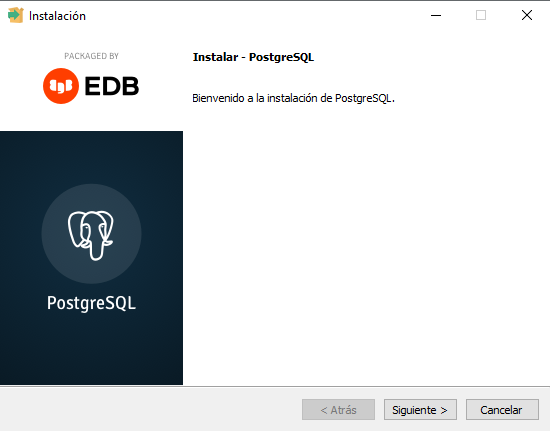
En el caso de que se desee utilizar el mismo SGBD para la práctica, en este caso Postgresql, la instalación es muy sencilla y simplemente hay que seguir los siguientes pasos:

1. Descarga postgresql: <https://www.postgresql.org/download/windows/>

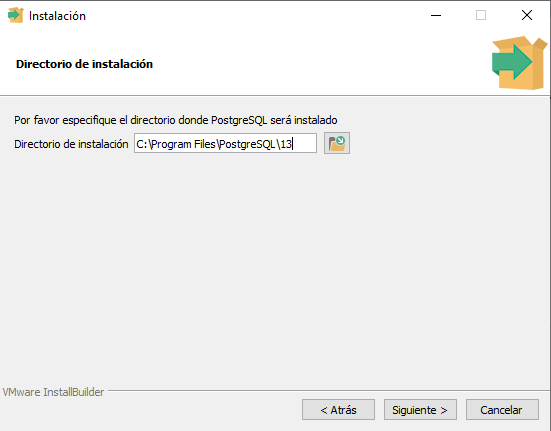


1. Seguir las instrucciones del instalador:

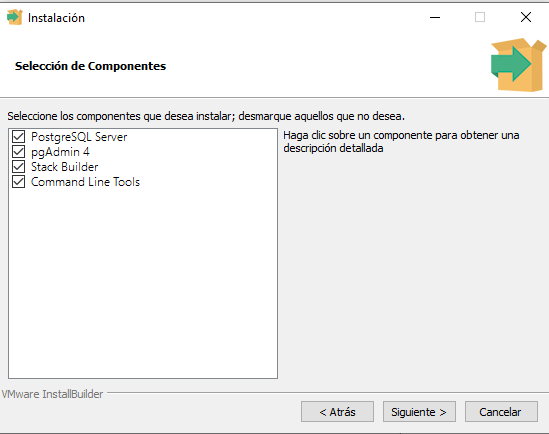
Paso 1



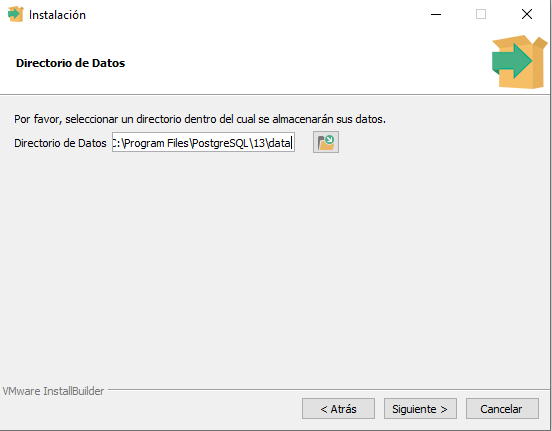
Paso 2



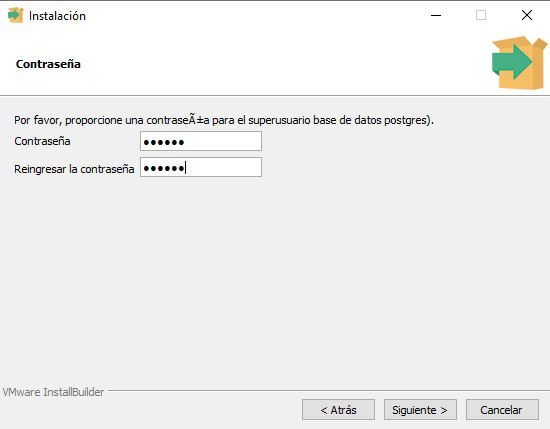
Paso 3: selección de componentes



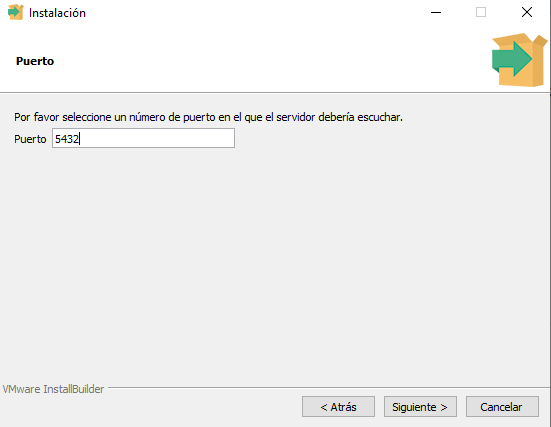
Paso 4: directorio donde almacenar los datos



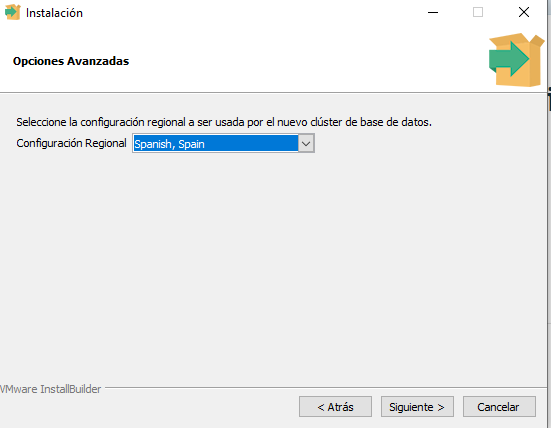
Paso 5: Introducir la contraseña para el superusuario postgres



Paso 6: seleccionamos el puerto



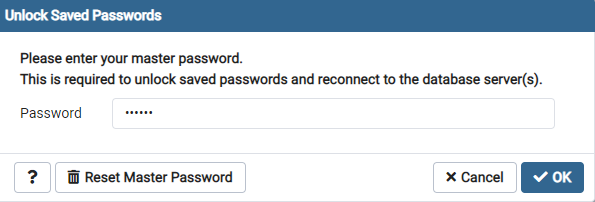
Paso 7: seleccionar configuración regional



Cuando finalice la instalación, abrimos pgAdmin:



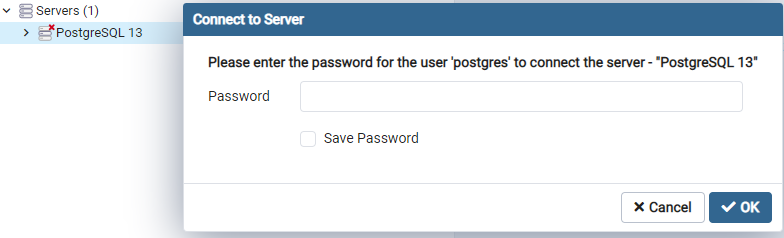
Al entrar nos pedirá el password que hemos puesto durante la instalación



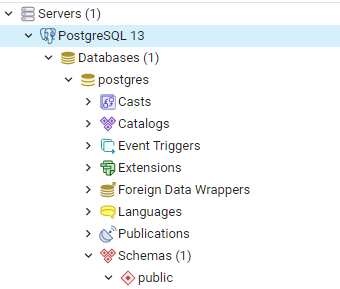
Si todo es correcto podríamos ver ya nuestro servidor



Nos conectamos para ver que todo es correcto:



Una vez identificados podremos ver nuestras bases de datos



## Dimensiones del modelo

Algo importante en el diseño de los procesos de aprovisionamiento de un DWH es la estrategia de carga de las dimensiones. En nuestro ejercicio vamos a contar con 2 dimensiones:

DIMCENTRO: con la información de los centros adheridos a *Traçacovid.*

DIMTIEMPO: con la información de tiempo.

En el caso de la DIMCENTRO se tratará de una SCD2 o dimensión lentamente cambiante. El término dimensiones lentamente cambiantes hace referencia a que los datos de las dimensiones cambian poco a poco a lo largo del tiempo. No son incrementales como los hechos. Cada día hay nuevas incidencias reportadas por los centros, sin embargo, la información de cada centro va cambiando poco a poco a lo largo del tiempo. Las dimensiones pueden ser estáticas como los meses, o cambiar muy lentamente (como el nombre de los centros), o incluso muy rápido como por ejemplo la edad de los alumnos. Hablar de estrategias SCD es hablar del modo de cargar dimensiones.

En este caso, la tabla incluirá los campos:

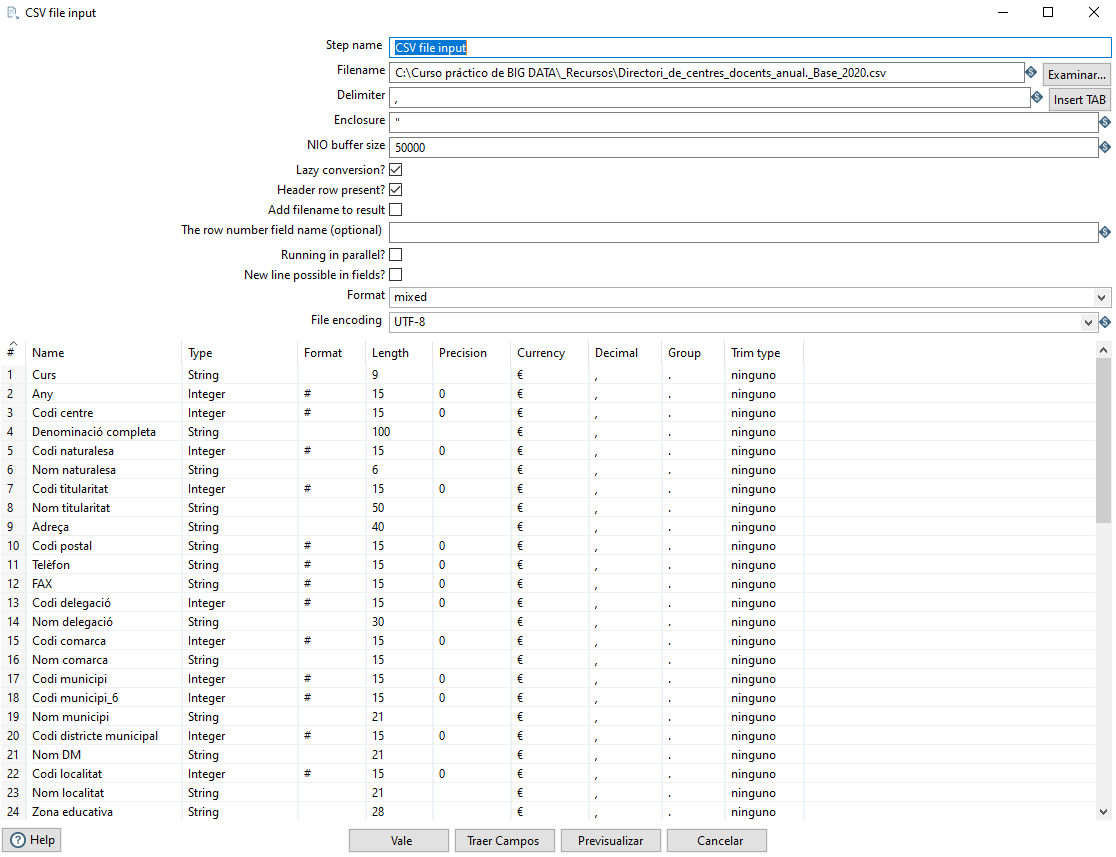
1. Fecha inicio de vigencia

2. Fecha fin de vigencia

Estas fechas nos permiten determinar en qué estado estaba la dimensión en cualquier fecha del calendario.

### CREACIÓN DE UNA TRANSFORMACIÓN ETL PARA CARGAR Y ADMINISTRAR LOS DATOS DE LOS CENTROS COMO UNA SCD TIPO 2

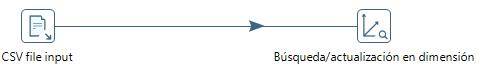
Para este tipo de Dimensiones, se creará un nuevo registro con el nuevo conjunto de registros. Se genera una nueva clave de sustitución, pero la clave nativa o natural se puede repetir junto con algunos de los otros valores de atributo. Pentaho Data Integration (Kettle) llama a este tipo de inserción “update” en el step Dimension Lookup/Update. Para ello, primero cargamos el fichero con los datos de incidencias a través de un CSV File Input:



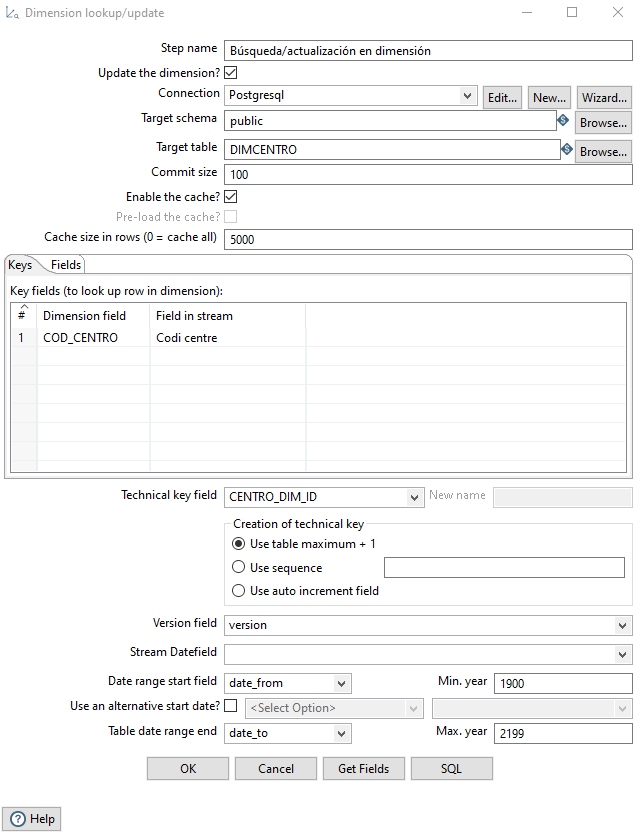
Luego añadimos un objeto de transformación Búsqueda/Actualización en dimensión:



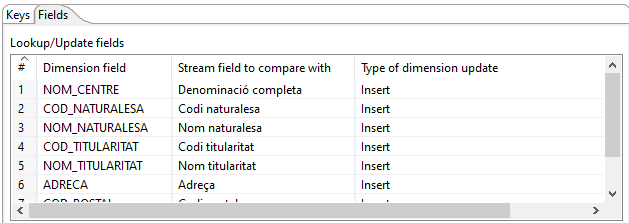
Unimos ambos objetos y damos a configurar este último:



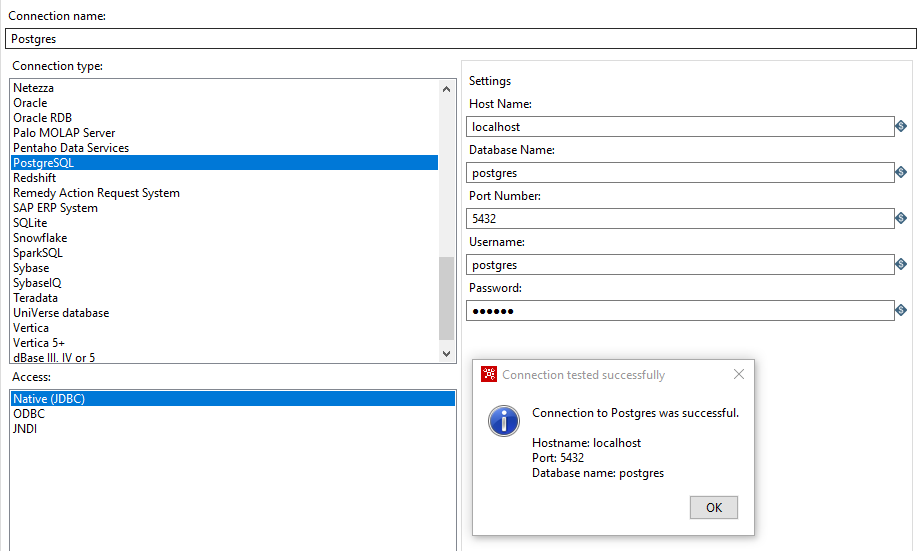
En la ventana de configuración indicamos la clave y los campos de versionado: DATE\_FROM, DATE\_TO, Y VERSION. Además, indicamos que la dimensión tendrá como clave técnica CENTRO\_DIM\_ID.



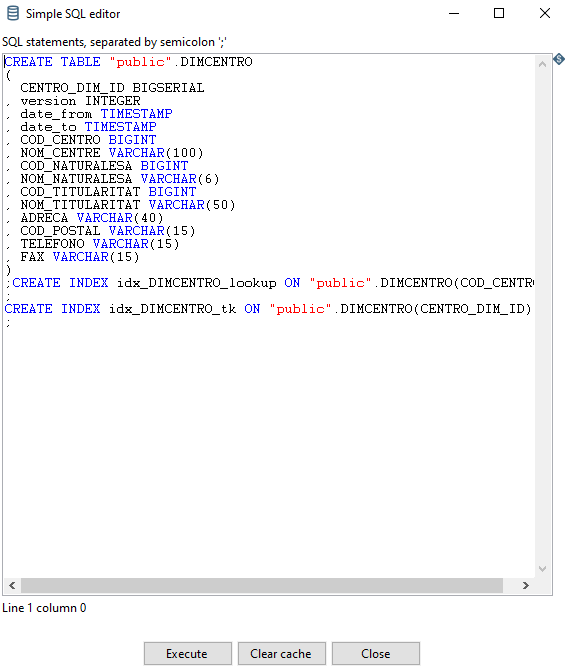
Configuramos el resto de campos en Fields:



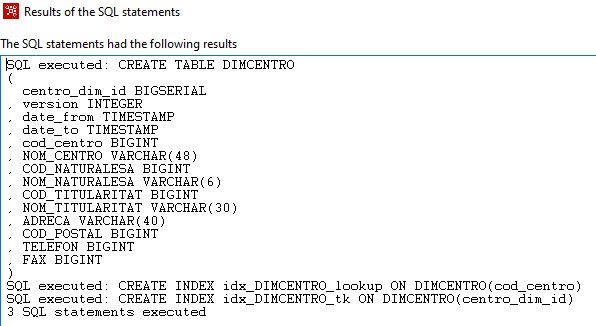
A continuación seleccionamos New Conection y la configuramos como se muestra para Postgres



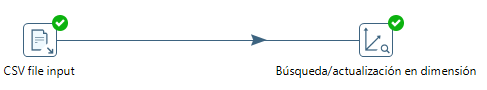
Una vez configurada la conexiónen el botón SQL le decimos que cree la nueva dimensión en nuestra base de datos, en el caso de esta práctica se ha utilizado postgresql.



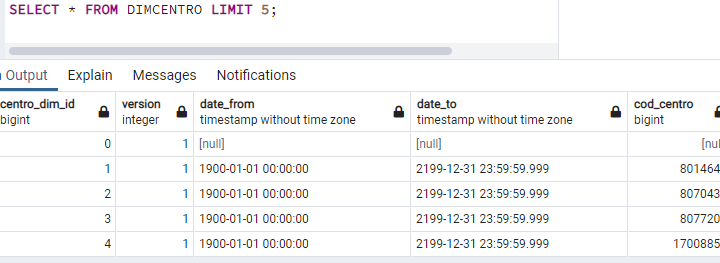
Si le damos a ejecutar nos crea automáticamente la nueva dimensión en la base de datos y esquemas indicados en la configuración devolviéndonos el resultado:



A continuación, lanzamos el proceso y comprobamos que se ejecuta correctamente:



Por último, miramos que los datos generados en la entidad destino sean correctos:



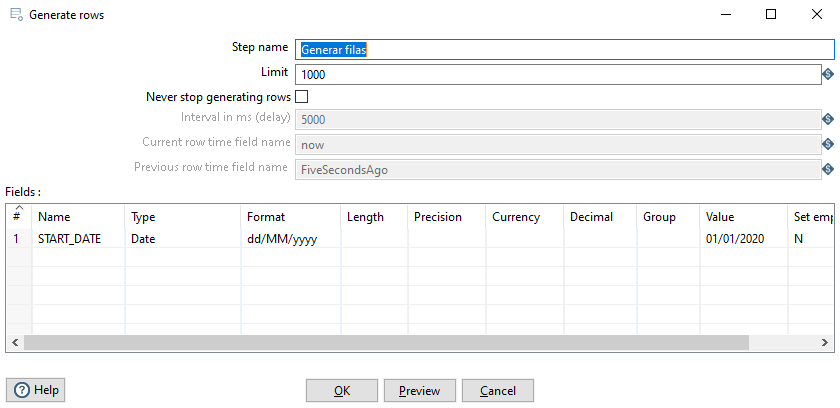
### GENERACIÓN DE UNA TABLA DE DIMENSIÓN DE FECHAS

Kettle tiene herramientas que pueden ser utilizadas para generar datos. Hay casos, como éste, en el que necesitamos generar datos para crear una dimensión. Las dimensiones de fecha y hora son claros ejemplos de esto. En este ejercicio, vamos a generar una dimensión de tiempo para utilizarla en nuestro Datawarehouse.

1. Crear una nueva transformación.
2. En la pestaña “Design”, abrir la “carpeta Input” y arrastra un elemento “Generate Rows” a la nueva transformación.



1. Doble click en el step “Generate Rows” para ver sus propiedades. Para crear la dimensión de Tiempo, se necesita generar una fila por cada día desde 01/01/2020. Para ello, asignar el valor de 1000 al parámetro “Limit”. En la tabla, campo “Fields”, crea un campo nuevo denominado “START\_DATE”. En la columna “Type”, selecciona “Date” y en “Format” escribe “dd/MM/yyyy”.
2. En la columna “Value” (Fecha de inicio) escribe: 01/01/2020



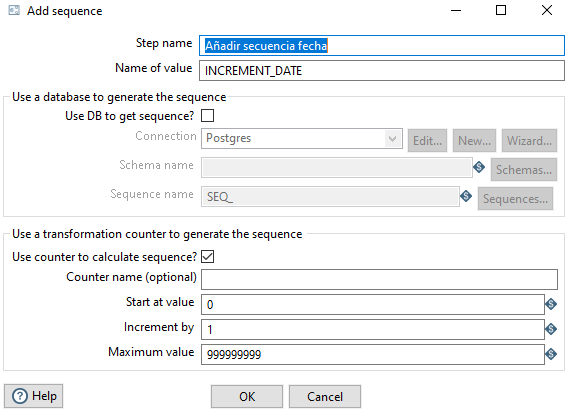
1. V. Haz clic en el botón “Preview” para ver los datos generados. Debes tener una lista de 1000 elementos con la misma fecha repetida una y otra vez. Cierra la Vista previa y haz clic en Aceptar.
2. VI. En la pestaña de diseño, abre la carpeta Transform y arrastra un step “Add sequence” a la transformación.



1. Conecta ambos steps. Establece la salida del “step Genertate Rows” como entrada del step “Add sequence”.



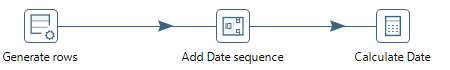
1. VII. Abre las propiedades del nuevo step. Cambia el nombre de este nuevo paso "Add date sequence" y proporciona un Name of value a “increment\_date”. Este paso generará un número incremental para cada registro de datos que reciba. El valor de inicio en el valor debe ser 0 y el incremento de debe establecerse en 1.



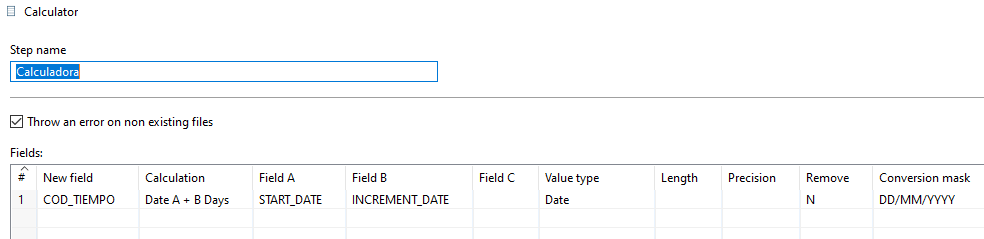
1. VIII. Abre la carpeta de “Transform” y, a continuación, arrastra y suelta un step “Calculator” en la transformación.



1. Enlaza el conector de salida del paso Add sequence a la entrada del step Calculator.

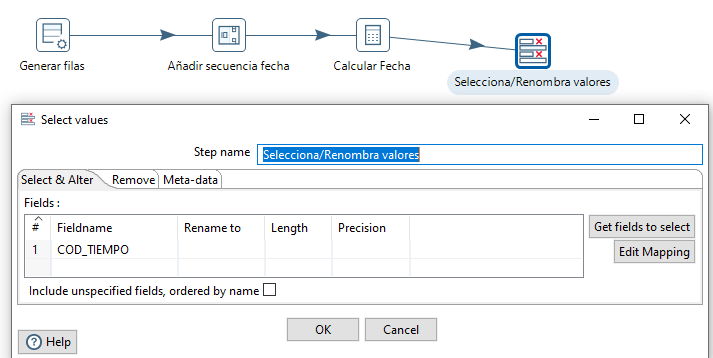


1. Abre las propiedades del nuevo step Calculator. Nombra el nuevo campo como "COD\_TIEMPO". Establece el cálculo en "Date A + B Days". Esto indica que tomaremos la fecha que llega al Campo A y le agregamos un número de días que llegarán al Campo B. Establece el campo A en START\_DATE y establezca el campo B en increment\_date. Establece el Value type en Date. Establece Remove en N y finalmente establece la Conversion mask en dd/MM/yyyy tal y como se muestra a continuación:

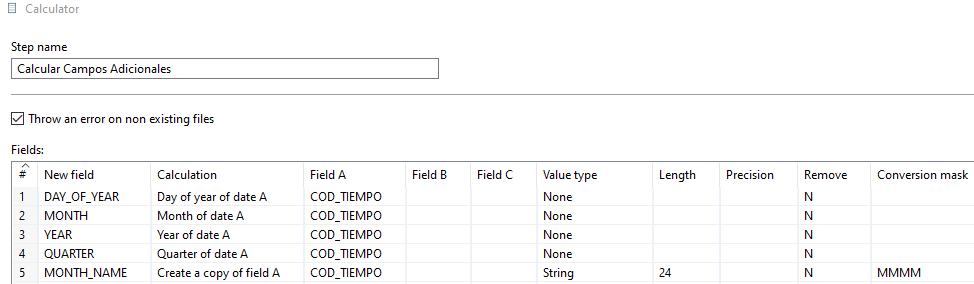


Click en el botón “OK” y cierra las propiedades del step. La transformación quedará ahora como se muestra a continuación:

1. En este punto, necesitamos SELECCIONAR SÓLO ALGUNOS CAMPOS de los que vienen a través del step “Calcular Fecha” para que tengamos únicamente el campo “COD\_TIEMPO”. Para ello, desde la carpeta Transform, arrastrar y soltar el step “Select values” en la transformación. Este paso sólo seleccionará el campo “COD\_TIEMPO” que se trasladará al siguiente paso de la transformación, cortando el paso al resto de los campos generados hasta ahora. Conecta la salida del step “Calcular Fecha” la entrada del paso “Select values”. Abre las propiedades del nuevo step “Select Values”. Selecciona "COD\_TIEMPO" como el nombre de campo como se muestra a continuación, haz clic en “OK” para cerrar el step.



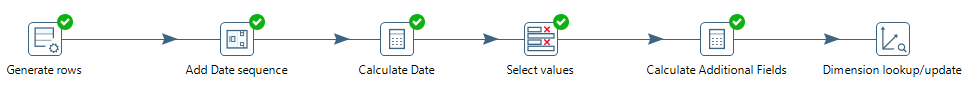
1. Ahora que tenemos una secuencia de fechas contiguas, podemos enriquecer la dimensión con campos calculados adicionales como el día de la semana, mes, año, etc. Agrega otro step “Calculator” a la transformación y vincula el conector de salida del step “Select values” a la entrada del segundo step “Calculator”. XIV. Abre las propiedades del step del nuevo “Calculator”. Cambia el nombre de este nuevo step “Calcular Campos Adicional”. Para cada nuevo campo, se establecerá “COD\_TIEMPO” para el campo A y los nuevos campos se calcularán todos a partir del “COD\_TIEMPO”. Introduce los nombres, cálculos y tipos de datos como se muestra en la siguiente tabla:



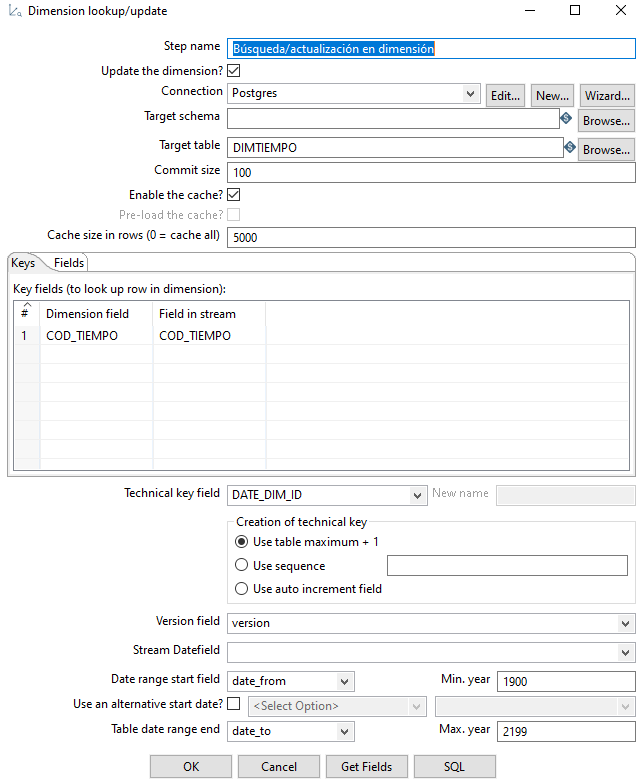
Para generar el nombre del mes necesitamos trabajar con una copia de “COD\_TIEMPO”, convertirla a una “String” y luego aplicar una máscara de “MMMM” que indica el nombre completo del mes.

1. Haz clic en OK para cerrar el step “Calcular Campos Adicionales”.
2. En este punto podemos probar la transformación y ver si son correctos los valores generados. Para ello, ejecutamos la transformación y hacemos click en la pestaña “Preview data” y seleccionamos último step. Si nos situamos en cada step, podemos comprobar la evolución de la transformación.
3. El paso final de la transformación es cargar estas fechas en la tabla de dimensiones de fecha denominada “DIMTIEMPO” junto con un conjunto de claves sustitutivas (técnicas) generadas.

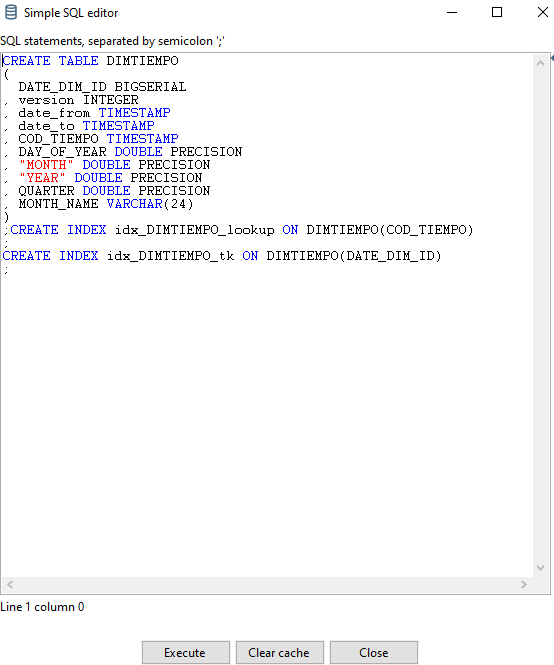
Arrastra y suelta un “step Dimension Lookup/Update” en la transformación. Conecta el conector de salida anterior del “Calculator” a este nuevo step.



1. Abre este nuevo paso y cambia el nombre a "Date Dimension Update" como se muestra a continuación. Utilice la misma conexión de base de datos de ejercicios anteriores. La nueva tabla se denominará DIMTIEMPO.
2. Haz clic en la pestaña “Fields” y, a continuación, haga clic en el botón “Get Fields”. Cambia el Tipo de actualización de la dimensión a “Update” para todos los campos restantes.



1. Revisa y ejecuta el script SQL create table para crear la tabla de dimensión de fecha en la base de datos de destino:



1. Cierra este último step “Dimension Lookup/Update”, guarda la transformación bajo el nombre ”Generate\_Date\_Transformation” y, a continuación, ejecútalo.
2. Revisa la tabla de la base de datos para asegurarse de que se han cargado todos los datos apropiados.

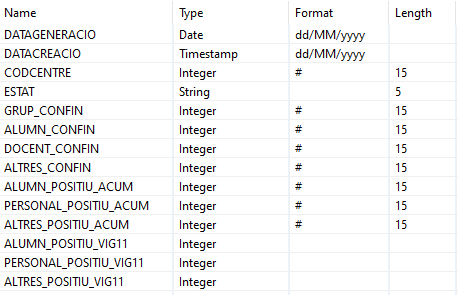
## Hecho del modelo

### DESARROLLO DE LA TRANSFORMACIÓN PARA CARGAR LA TABLA DE HECHOS

En esta sección abordaremos los pasos para crear la transformación ETL que construirá la tabla de hechos HECINDICACIONCOVID.

A partir de los datos de “Dades Covid”, se especificarán búsquedas en las dimensiones para centros y fecha, para a continuación, trasladar el conjunto de claves sustitutivas a la tabla de hechos junto con el acumulado de alumnos confinados, que será la única métrica.

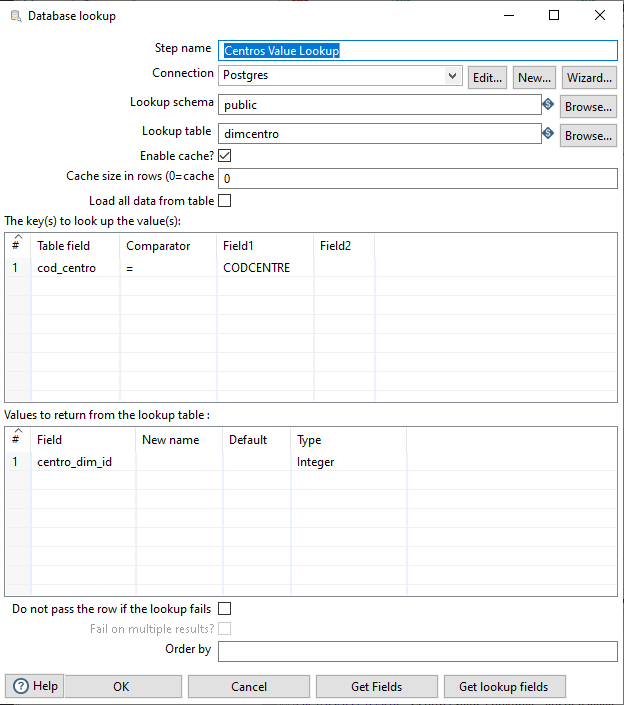
El fichero de entrada para esta transformación es “Dades\_COVID-19\_als\_centres\_educatius.csv”, que contiene las siguientes columnas:



1. Se comienza creando una nueva transformación. Arrastra un step de entrada de archivo CSV en la transformación. Nombra este step como “Dades Covid”. Busca y seleccione el archivo Dades\_COVID-19\_als\_centres\_educatius.csv.csv. Por último, hacer clic en el botón “Get Fields” para introducir todos los campos. Importante, asegúrate que el campo Lazy Conversation no está chequeado.
2. Arrastra y suelta un nuevo step de “Database Lookup” a la transformación. Vincula el conector de salida del paso de entrada CSV de Dades Covid a este nuevo step “Database Value Lookup”.

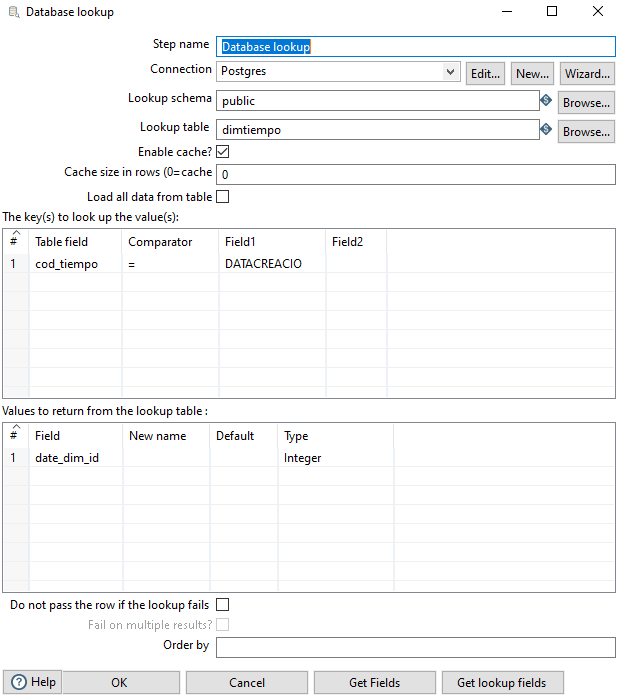
Configura el paso “Database Value Lookup”:

* Nombre: "Centros BD lookup".
* Este paso se utilizará para buscar la clave de centro “cod\_centro” para encontrar la clave de sustitución de dimensión de centros.
* Utiliza la misma conexión a base de datos que se ha utilizado para crear la dimensión de Centros.
* Especifica “cod\_centro” = “CODCENTRE” en el campo Lookup value.
* Para el apartado “Values to return from the lookup table”, introduce la clave técnica, que en este caso es “centro\_dim\_id”.

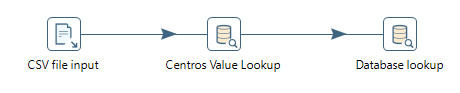


**Importante**: Asegurar que la casilla “Enable cache” está activa, de lo contrario, los datos de dimensión que se han actualizado pueden no estar representados en los datos de búsqueda.

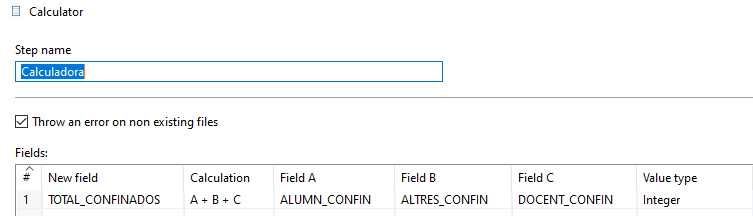
1. Procedemos de manera análoga con la dimensión de fecha:



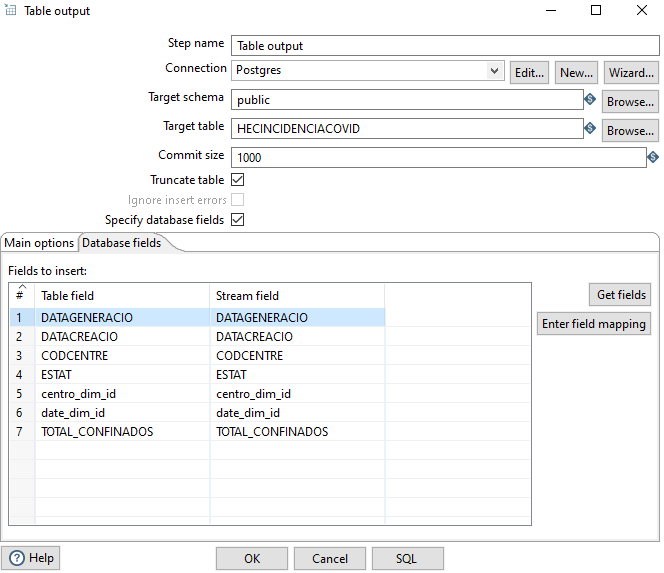
En este punto, la transformación debería tener el siguiente aspecto:

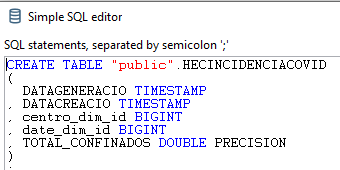


1. Realiza una vista previa de la transformación. Comprueba que no hay errores y observa la salida de datos. Deberías ver que, además de los valores de clave originales, se han agregado las claves de sustitución (técnicas) para Centro y Tiempo para cada registro de datos.
2. Ahora que todos los pasos de búsqueda de dimensión se han añadido, podemos hacer un cálculo sencillo, por ejemplo, sumando todos los confinados: alumnos, docentes y personal externo. Arrastra un step “Calculator” en la transformación. Establece el conector de salida del paso Date “Database Value Lookup” a la entrada del nuevo step “Calculator”. Cree un nuevo campo denominado "TOTAL\_CONFINADOS" con cálculo "A + B + C".

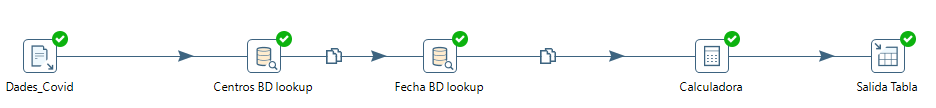


1. Por último, añade un step “Table output”. Utiliza la misma conexión a base de datos de destino que se ha utilizado para crear las dimensiones anteriormente. Especifica HECINCIDENCIACOVID como la Target table. Si esta tabla no existe se crea mediante el botón SQL

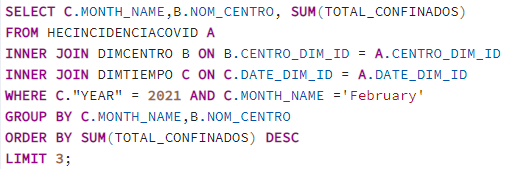




1. Ejecuta la transformación.



**EJERCICIO**: Lanza la siguiente sentencia SQL sobre la bbdd y adjunta el resultado obtenido en la casilla dispuesta a continuación:



|  |
| --- |
|  |

Para hacer un seguimiento del curso enviar esta memoria con el resultado del ejercicio anterior a [clopez@teralco.com](mailto:clopez@teralco.com)