Práctica-3: Herramientas ETL PDI (Pentaho Data Integration).

UNIVERSIDAD DE GRANADA

Sistemas Multidimensionales (2019-2020)

Daniel Bolaños Martínez
danibolanos@correo.ugr.es
Grupo 2 - Viernes 15:30h

Índice

1. **Crea una BD PosgreSQL (cordoba_danibolanos). En el esquema public de esa BD crea las tablas cuando, donde y padron. La estructura de estas tablas ha de ser similar a la de las hojas correspondientes del archivo cordoba-ETL-danibolanos.xlsx de la práctica anterior. Crea las transformaciones y jobs como se especifica en el guión.**

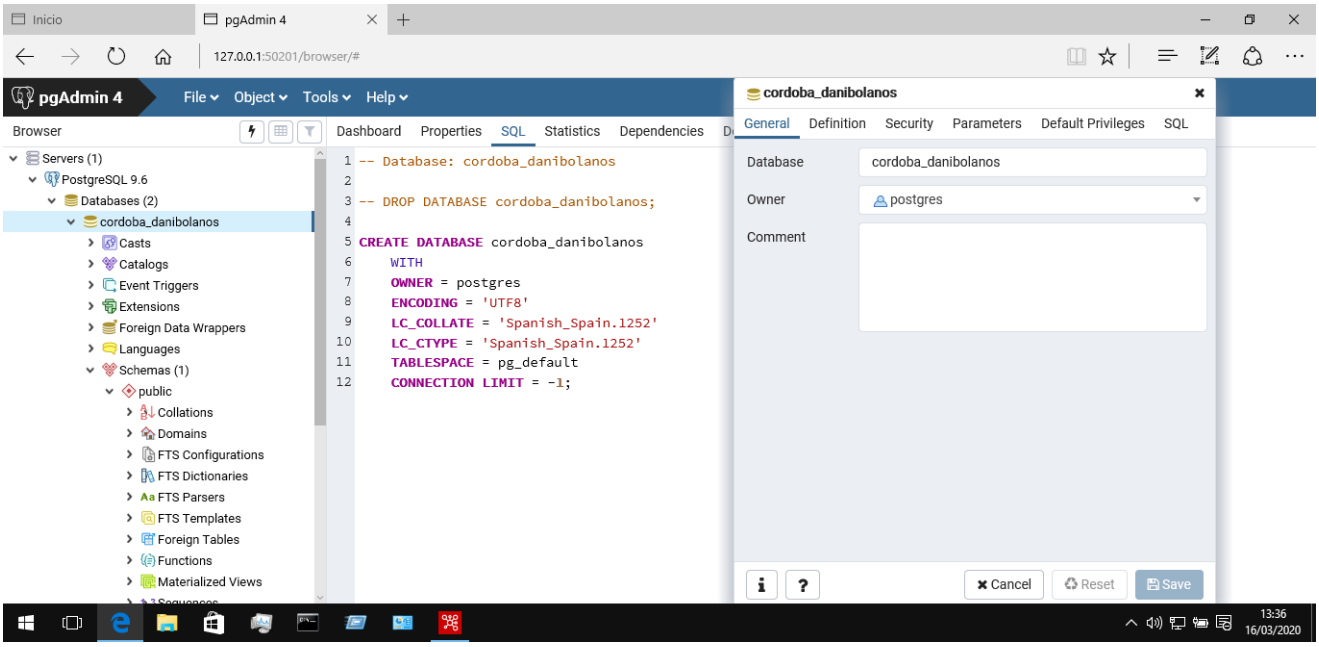

Figura 1: Creación de la BD **cordoba_danibolanos** en *pgAdmin*.

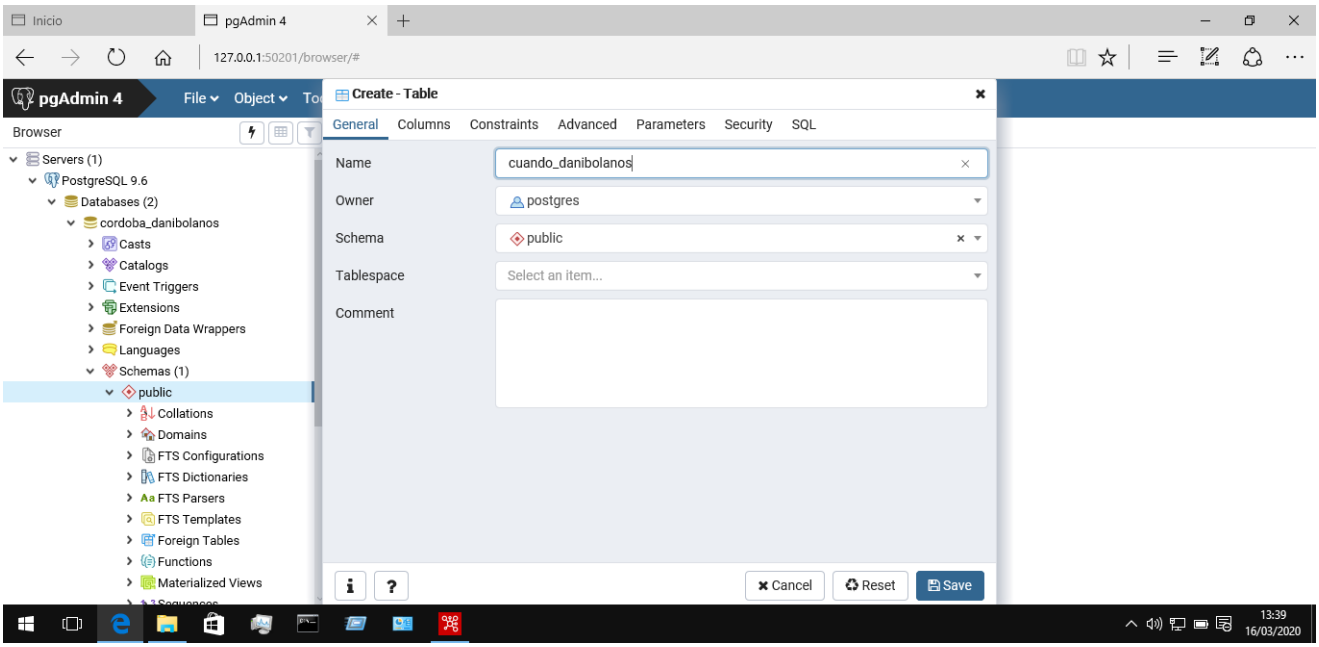

Figura 2: Creación de las tablas **cuando/donde/padron_danibolanos**.

Creamos las tablas **cuando_danibolanos, donde_danibolanos** y **padron_danibolanos** para la base de datos **cordoba_danibolanos** creada anteriormente.

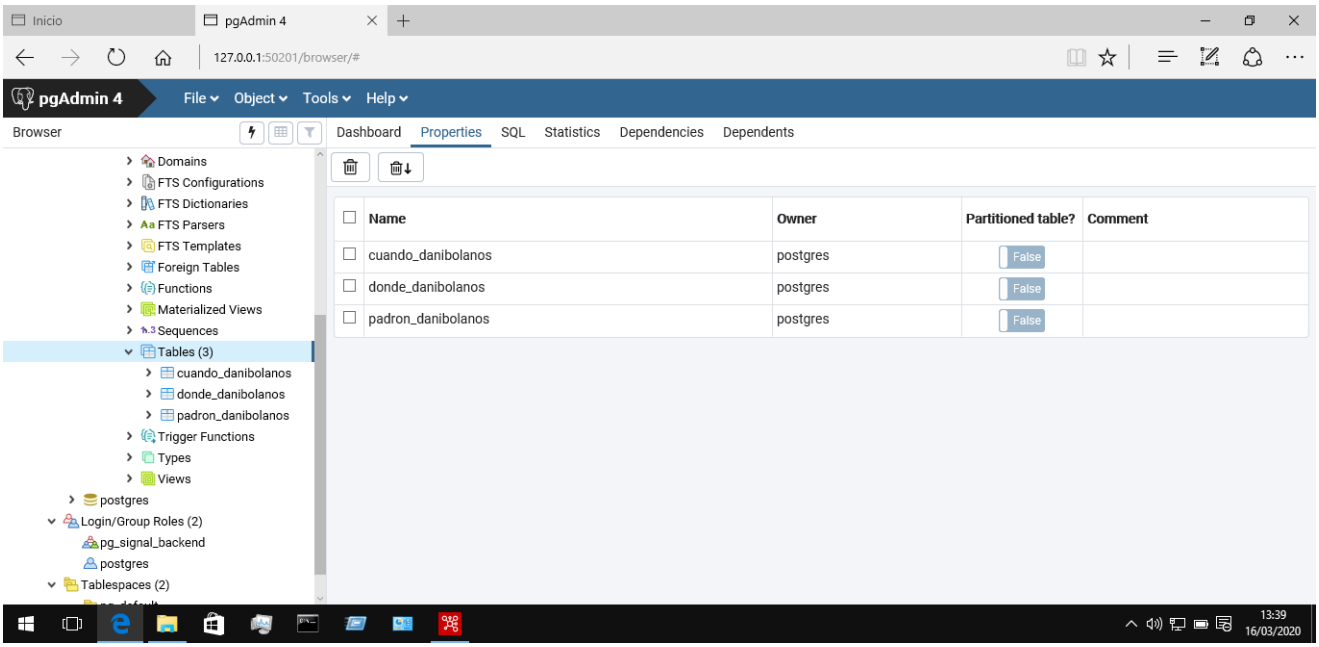

Figura 3: Tablas creadas en la BD.

Podemos observar las tres tablas creadas para la base de datos **cordoba_danibolanos**.

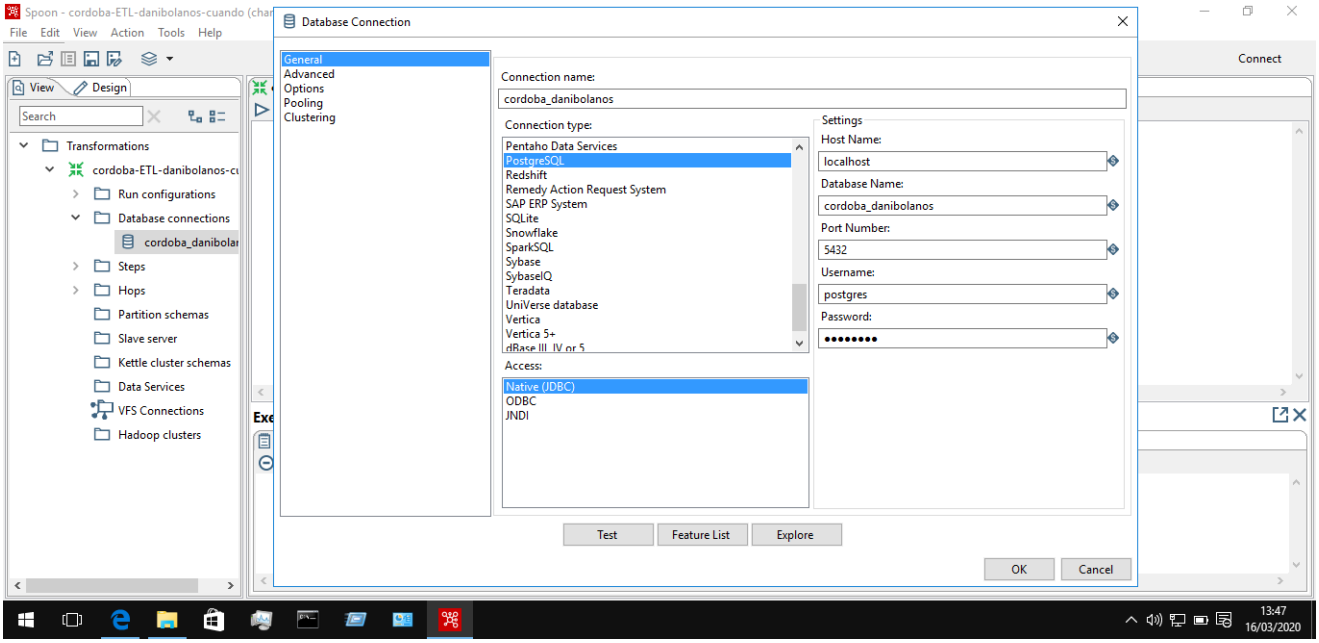

Figura 4: Creamos la conexión de la base de datos en *Spoon*.

Configuramos la conexión a la BD que hemos creados en *pgAdmin* para las transformaciones en *Spoon*. Usamos la configuración especificada en el guión.

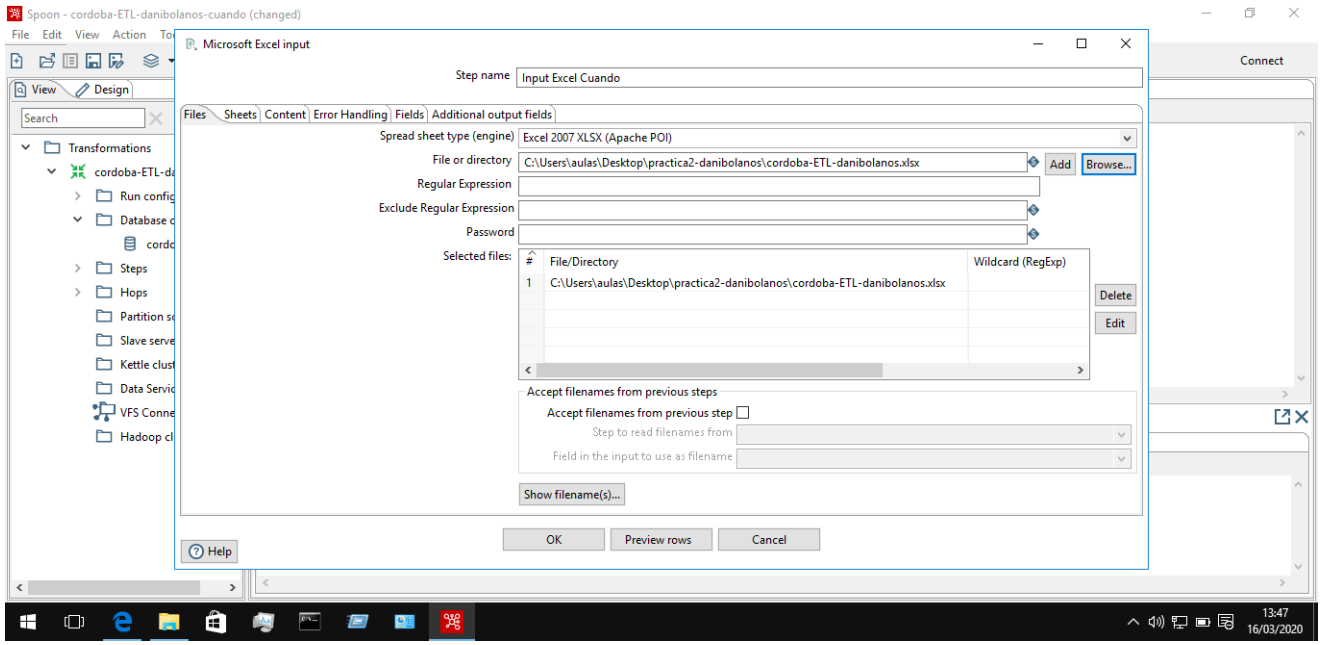

Figura 5: Extraemos los datos con el paso *Input Excel*.

Configuramos el paso *Input Excel* la obtención de los datos del archivo **cordoba-ETL-danibolanos.xlsx**.

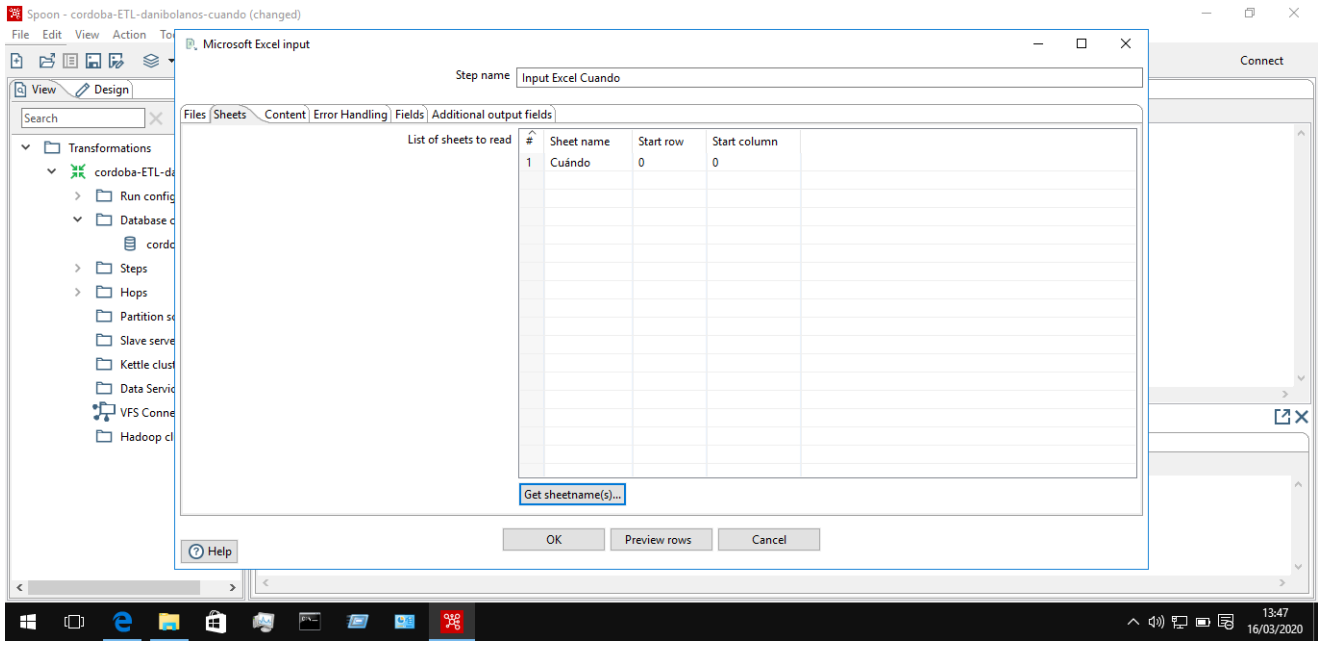

Figura 6: Extraemos los datos de la hoja **Cuándo**.

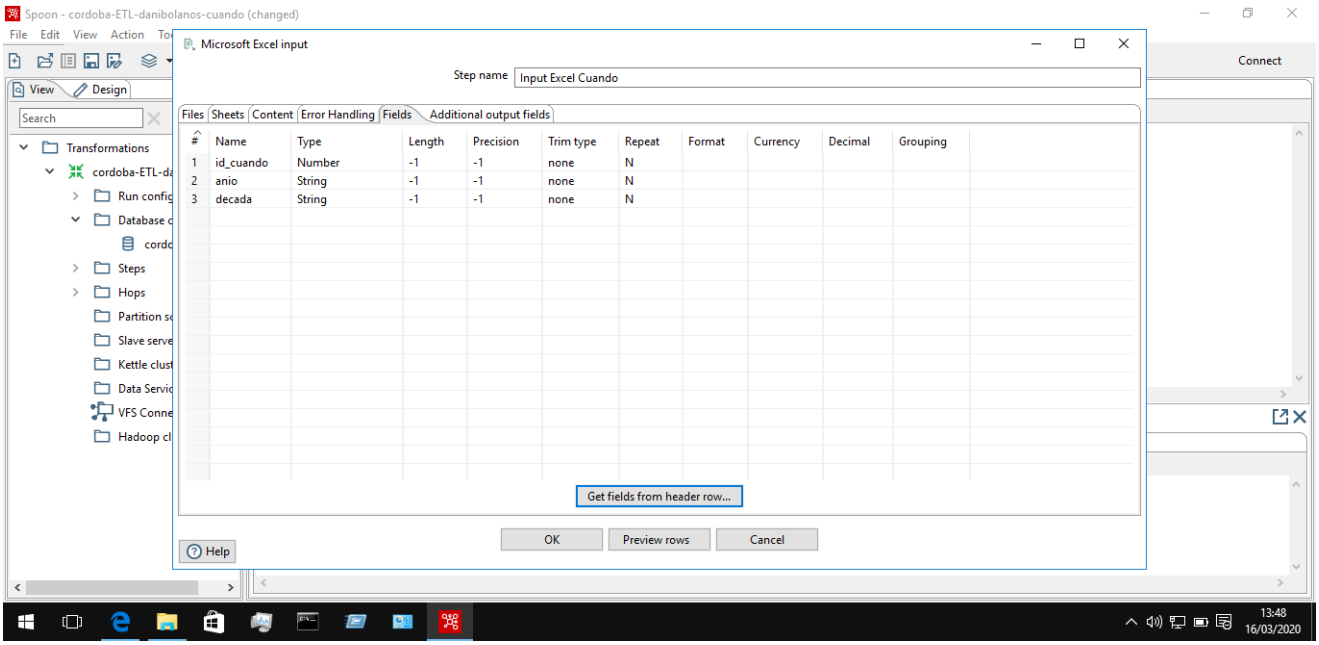

Figura 7: Campos de la hoja **Cuándo** con los criterios especificados.

Para cada hoja (**Cuándo, Dónde, Padrón**) del archivo excel generado en la práctica anterior, seleccionamos los campos y ponemos los nombres en minúscula, sin tildes y utilizando el criterio *camel_case*.

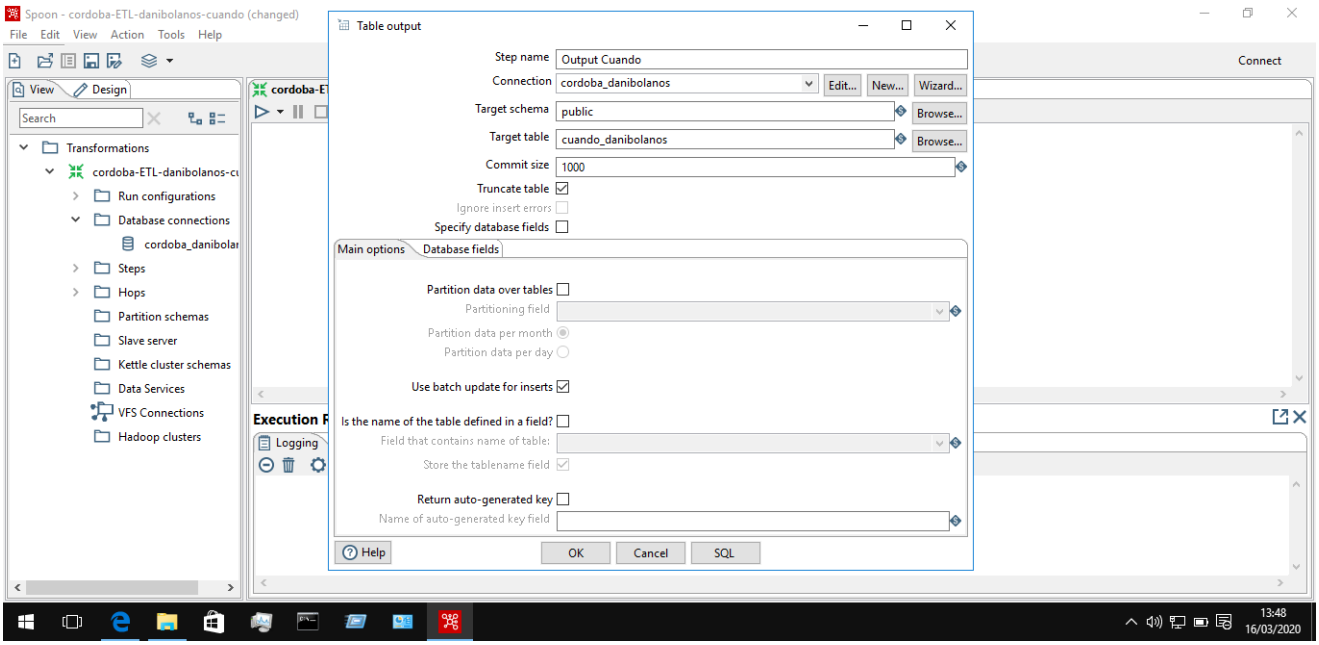

Figura 8: Configuración del paso *Table Output*.

Configuramos tal y como se indica en el guión para la hoja **Cuándo** el paso *Table Output* que se encargará de crear la tabla con los datos leídos para añadirla a la BD **cordoba_danibolanos** de *pgAdmin*.

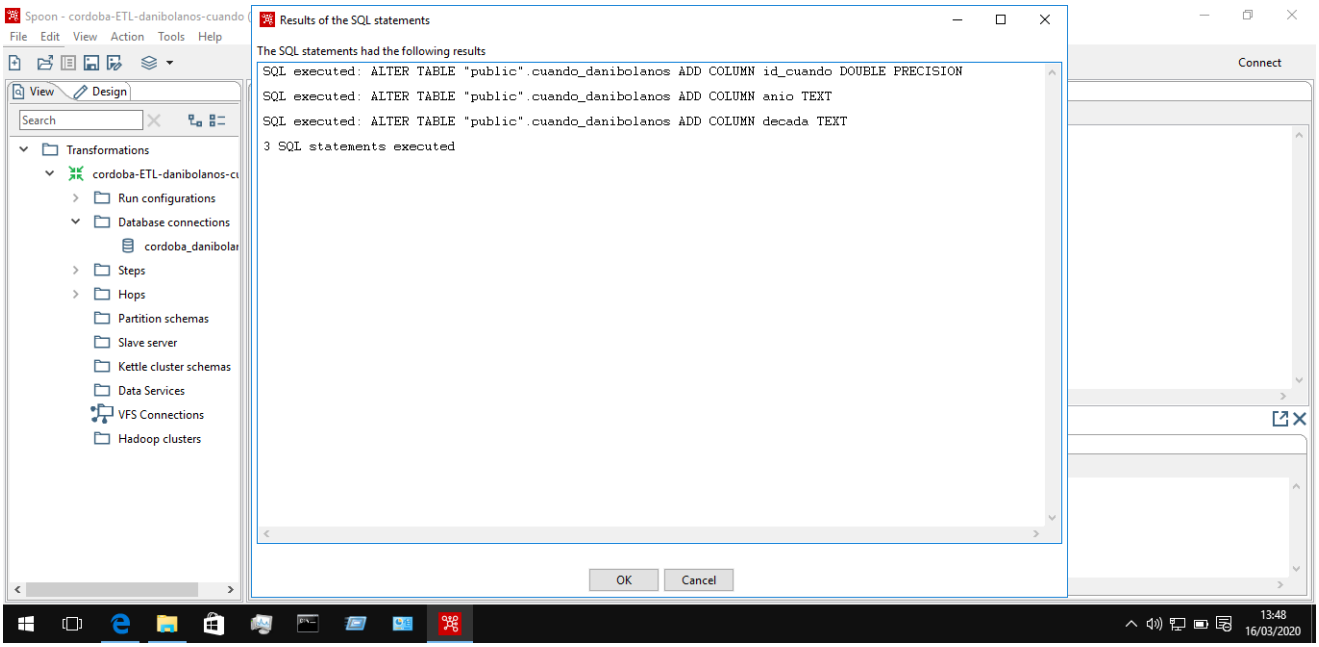

Figura 9: Generamos las sentencias SQL.

Al pulsar sobre el botón *SQL* del paso Table Output generamos las sentencias *SQL* encargadas de crear las tablas en la BD.

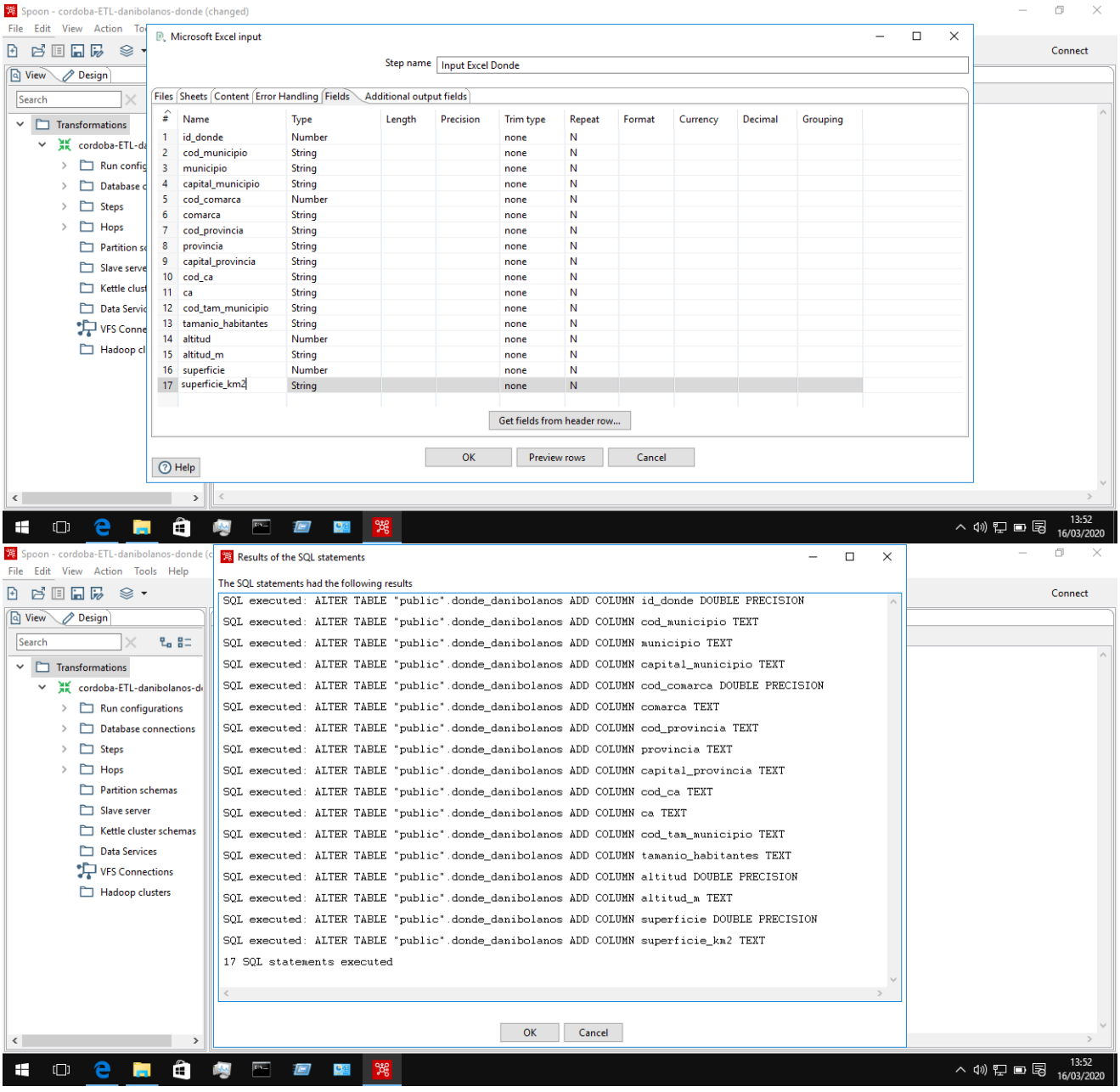

Figura 10: Realizamos el mismo proceso para **Dónde**.

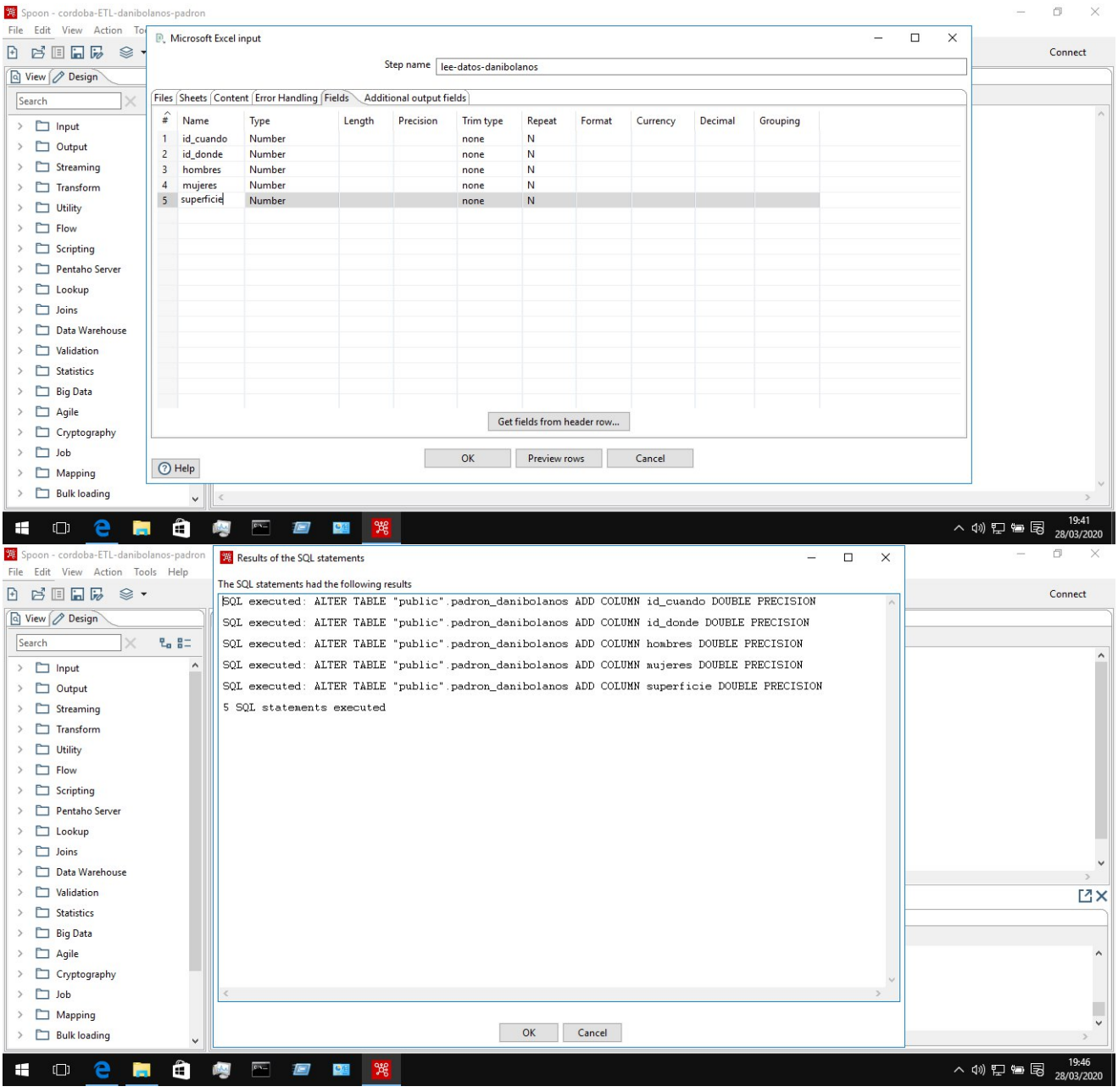

Figura 11: Realizamos el mismo proceso para **Padrón**.

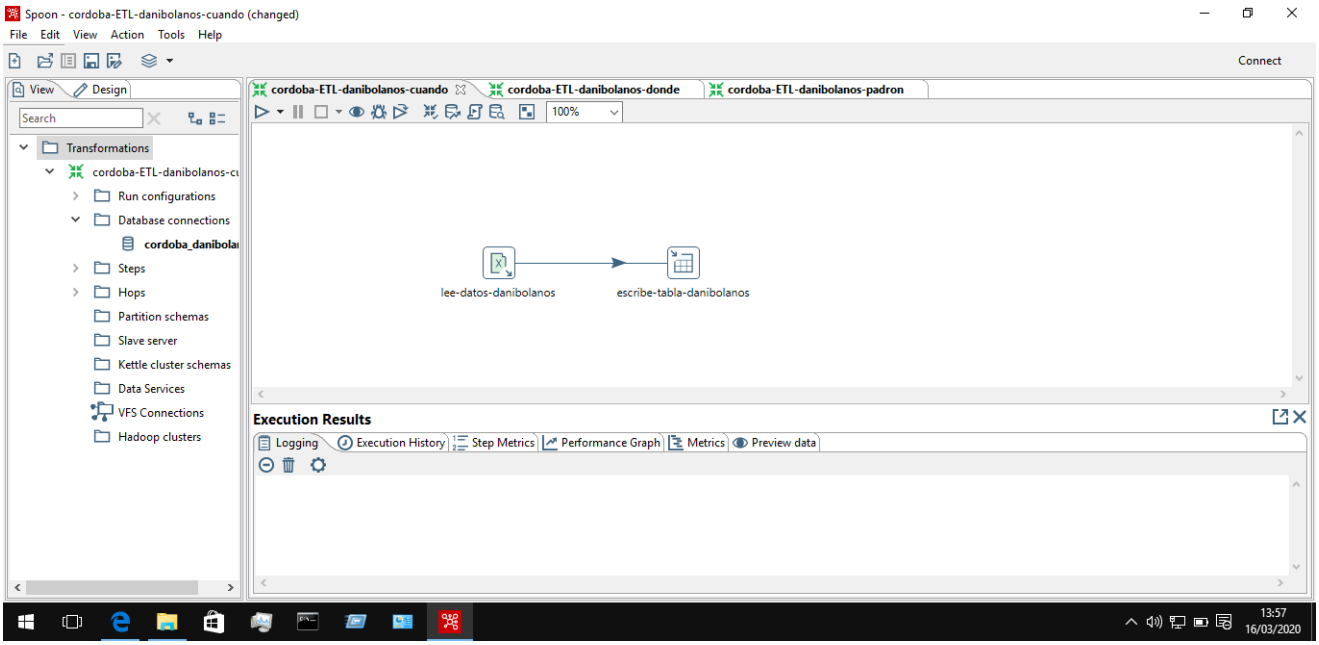

Figura 12: Ejemplo de la transformación **cordoba-ETL-danibolanos-cuando**.

Para las hojas **Dónde** y **Padrón** se han creado las transformaciones (**cordoba-ETL-danibolanos-donde** y **cordoba-ETL-danibolanos-padron**) de forma similar a la explicada.

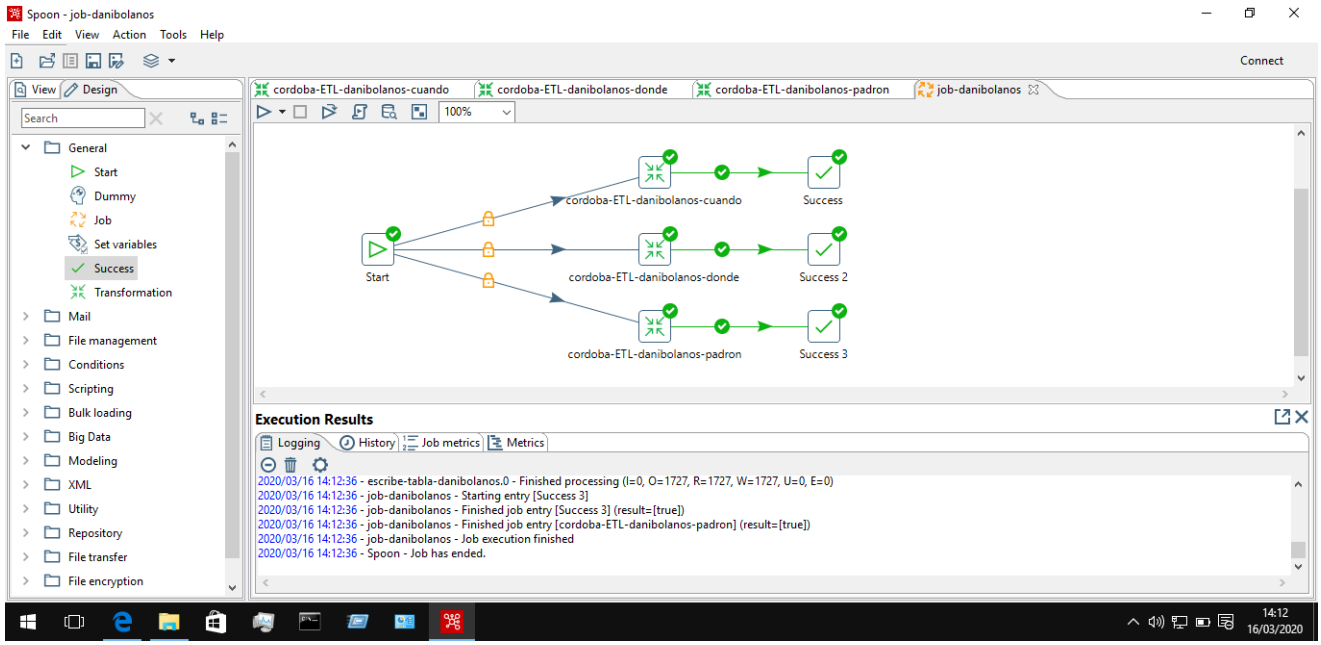

Figura 13: Job generado para las transformaciones.

Creamos una tarea 'Job' para ejecutar las tres transformaciones encargadas de crear las tablas en la BD para las tres tablas especificadas.

Creamos un paso **Start** para empezar el 'Job' y utilizaremos un paso **Sucess** para cada transformación para controlar si funciona todo bien o si se generan errores.

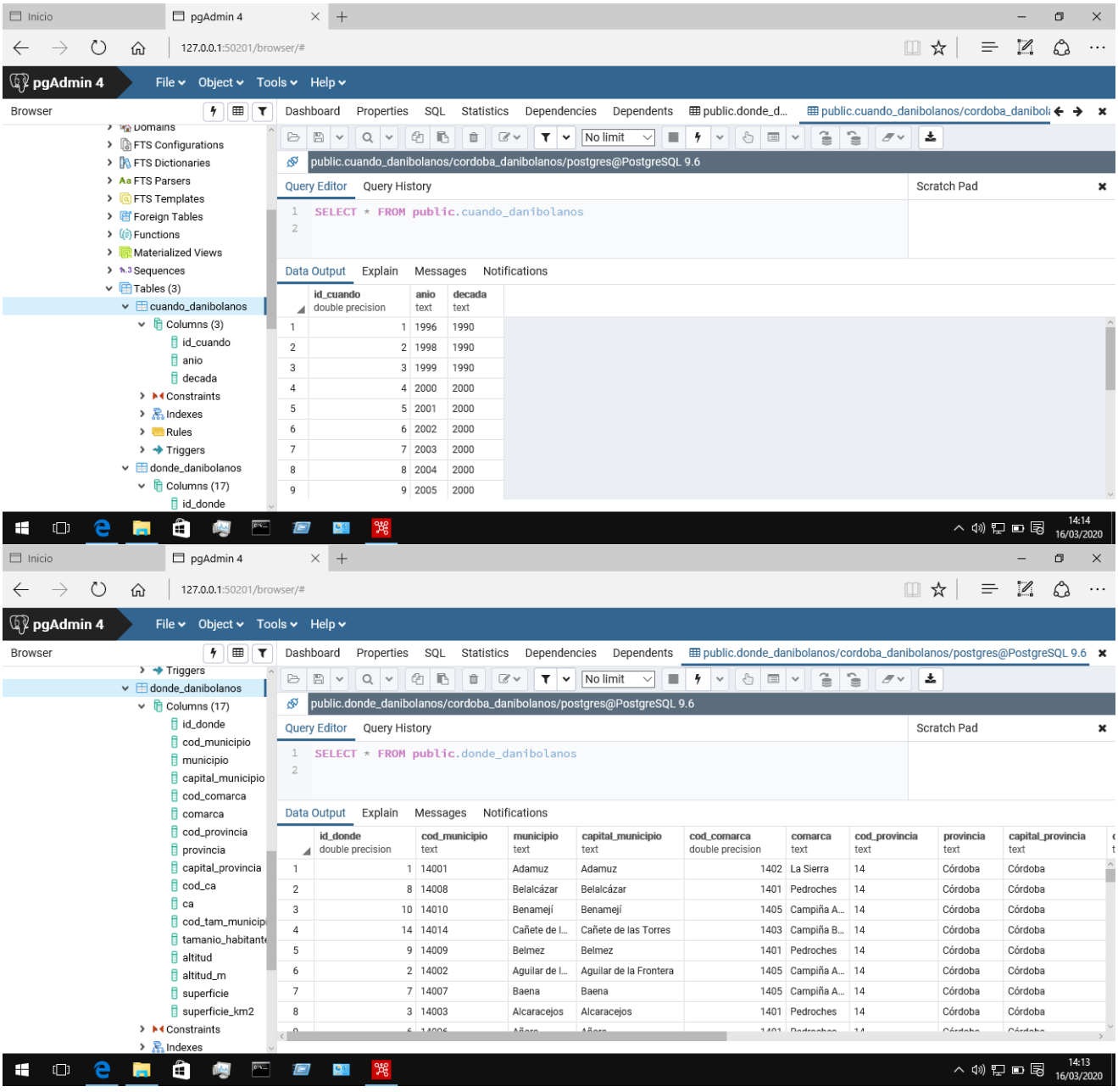

Figura 14: Datos para tablas **cuando** y **donde** en *pgAdmin*.

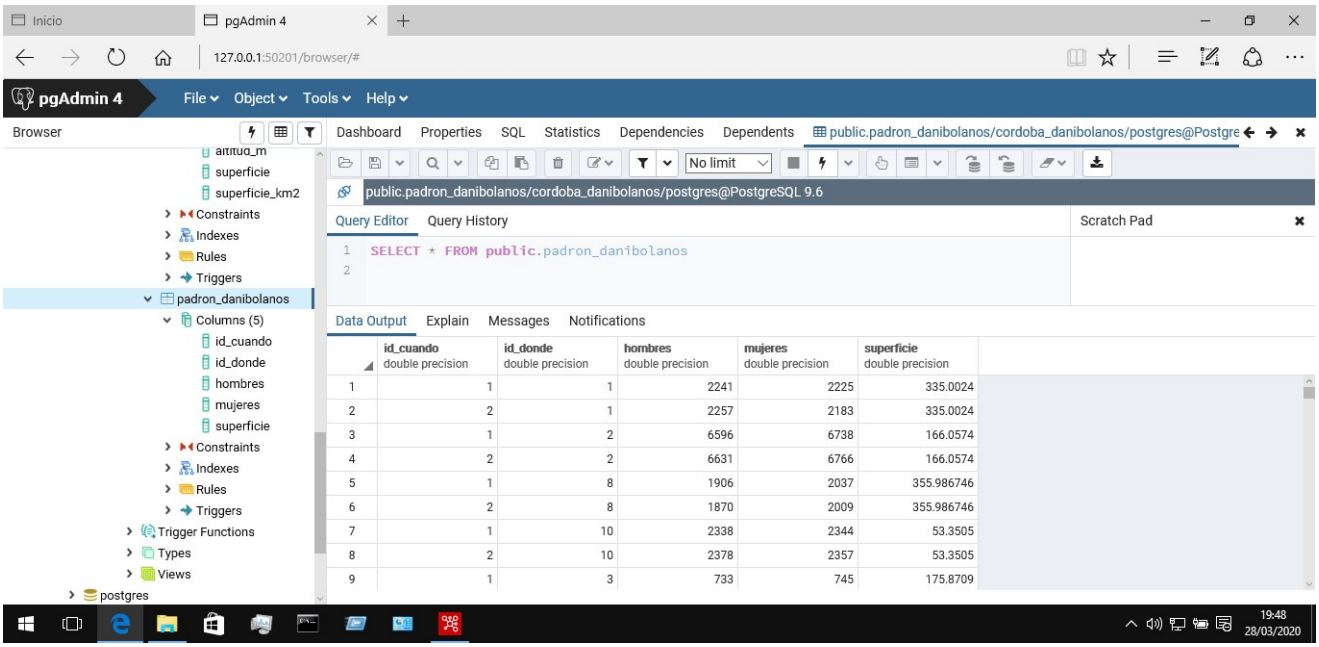

Figura 15: Datos para tabla **padron** en *pgAdmin*.

Podemos observar como se han añadido correctamente los datos de las tres hojas a la BD que habíamos creado previamente en *pgAdmin*.

2. Crea una BD PosgreSQL (prueba). En el esquema public de esa BD crea la cuando_danibolanos. La estructura de estas tablas ha de ser similar a la de las hojas correspondientes del archivo cordoba-ETL-danibolanos.xlsx de la práctica anterior. Define el contenido de esa tabla mediante una transformación usando como origen la hoja Provincia del archivo generado mediante Power Query en la actividad anterior.

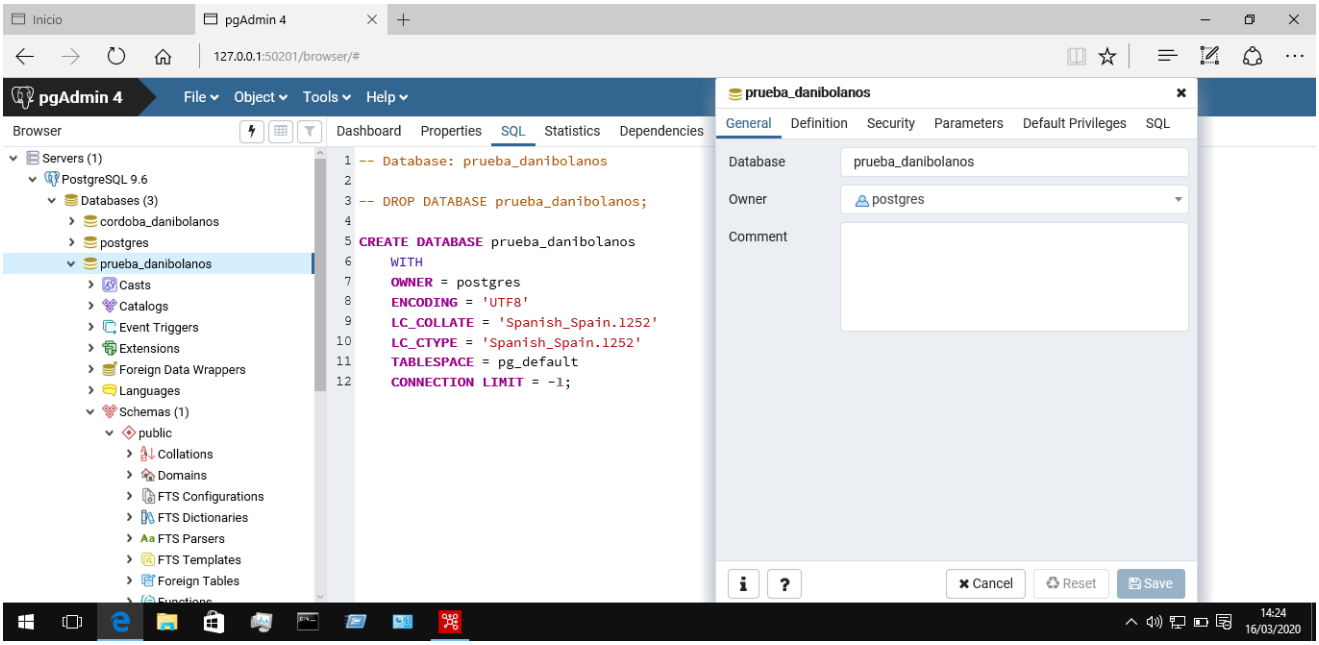

Figura 16: Creamos la BD **prueba_danibolanos** en *pgAdmin*.

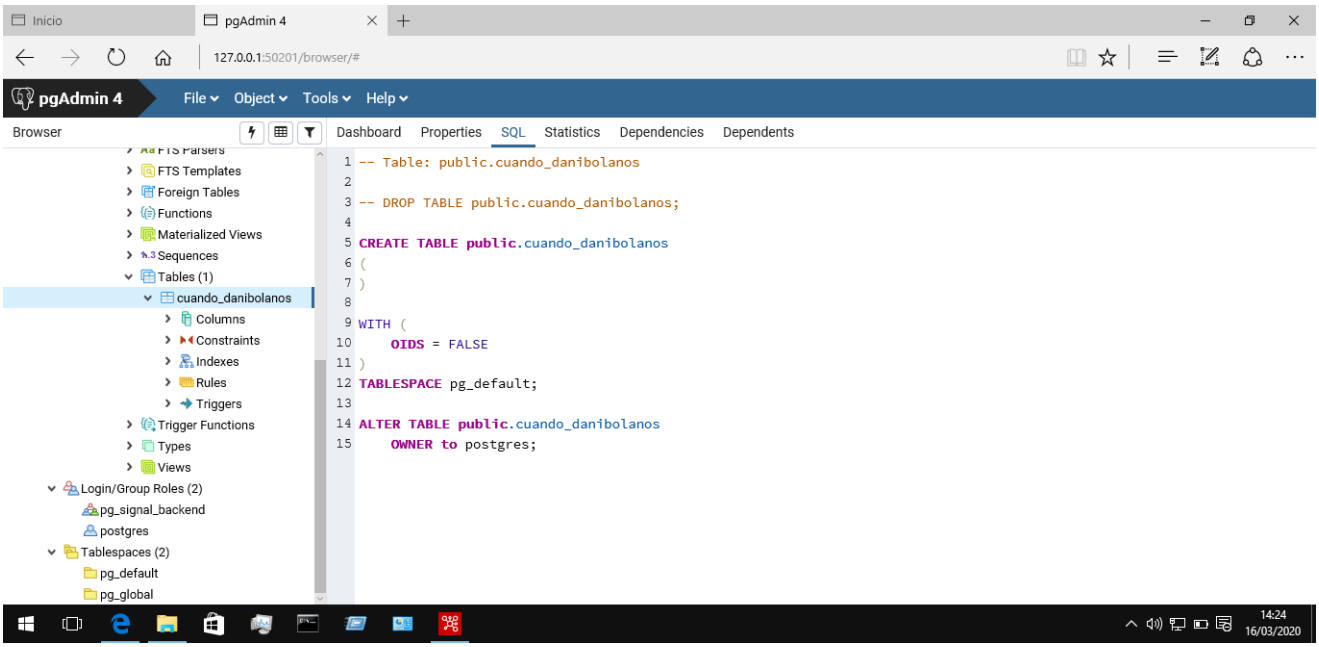

Figura 17: Creamos la tabla **cuando_danibolanos** en la BD.

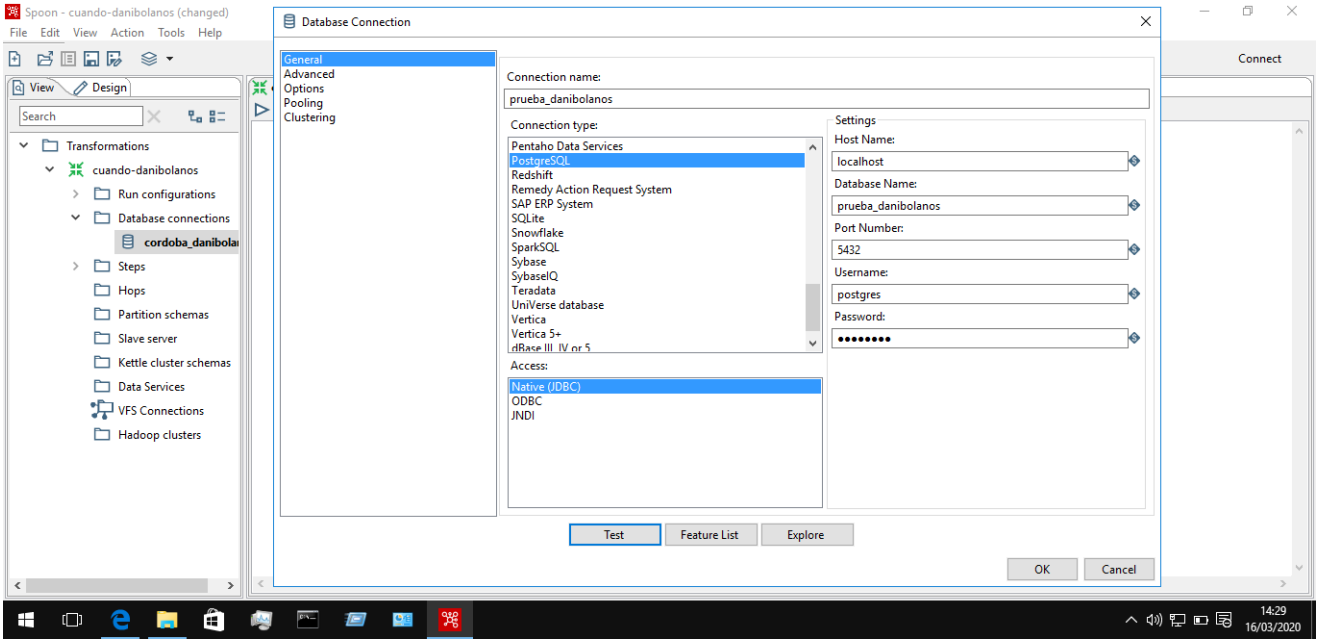

Figura 18: Configuramos la conexión para la BD **prueba_danibolanos**

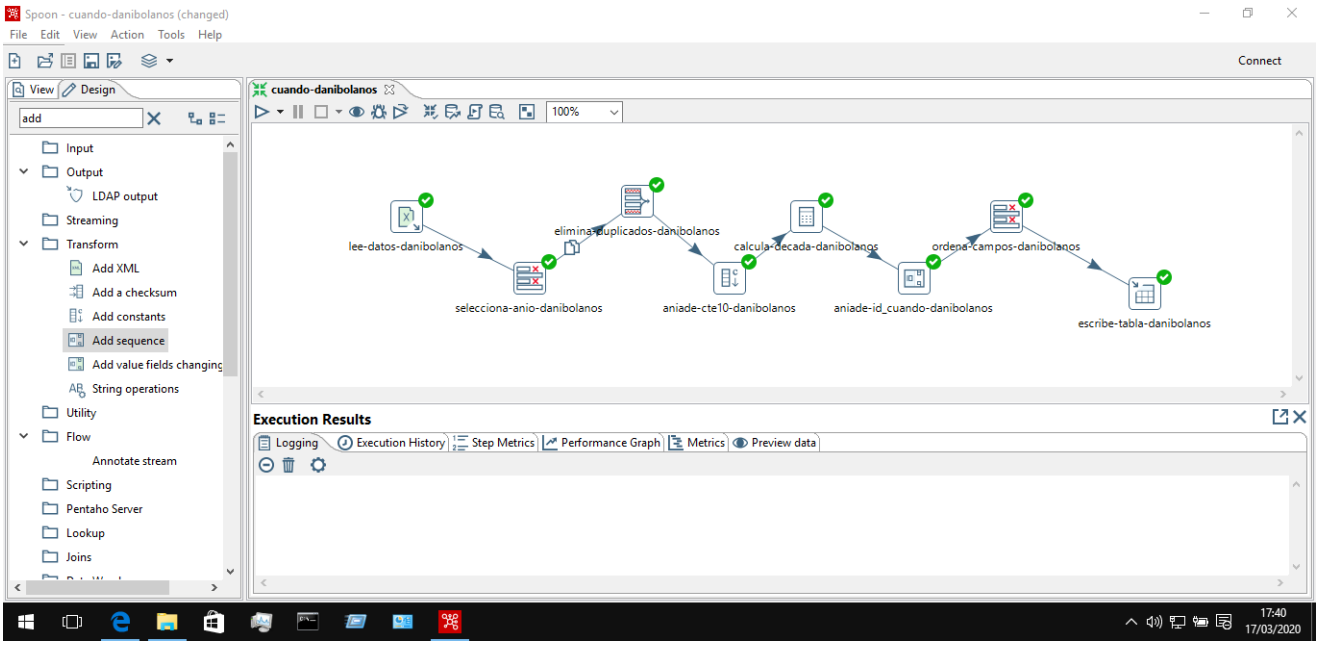

Figura 19: Configuramos los diferentes pasos en la transformación.

Para la transformación **cuando-danibolanos** creamos los siguientes pasos para pasar de la hoja **Provincia** a la tabla **cuando**.

Para ello necesitamos quedarnos sólamente con el campo **anio** y completar la tabla con los campos **id_cuando** y **decada** a partir de la información que aporta el año.

De izquierda a derecha se han usado los pasos: *Excel input file, Select values, Unique rows, Add constants, Calculator, Add sequence, Select values* y *Table output*.

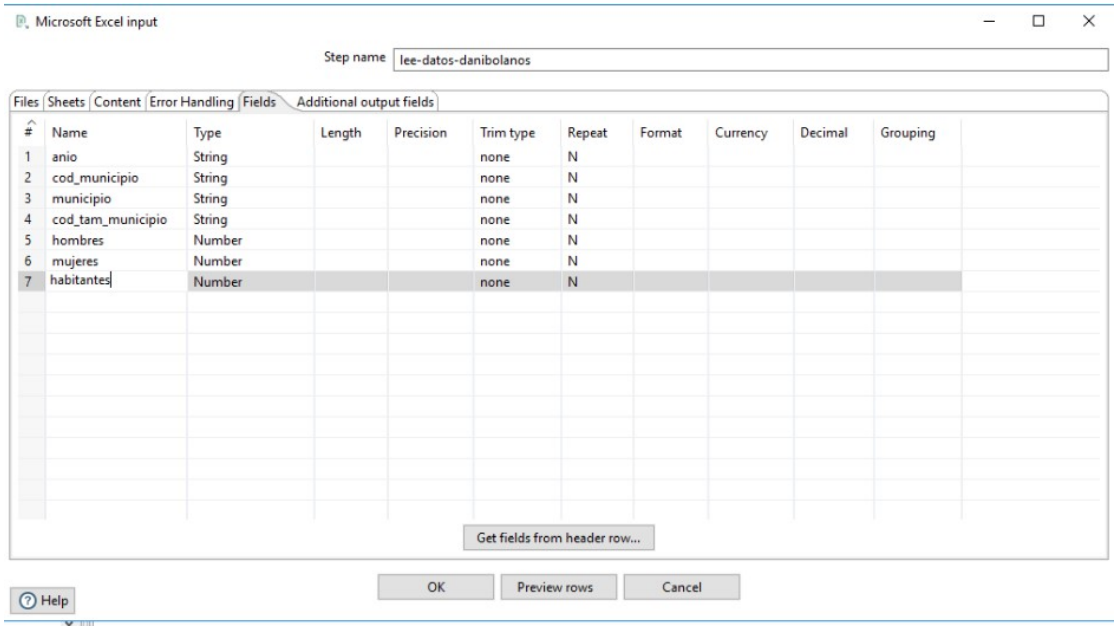

Figura 20: Leemos los datos de **Provincia** con criterio *camel_case*.

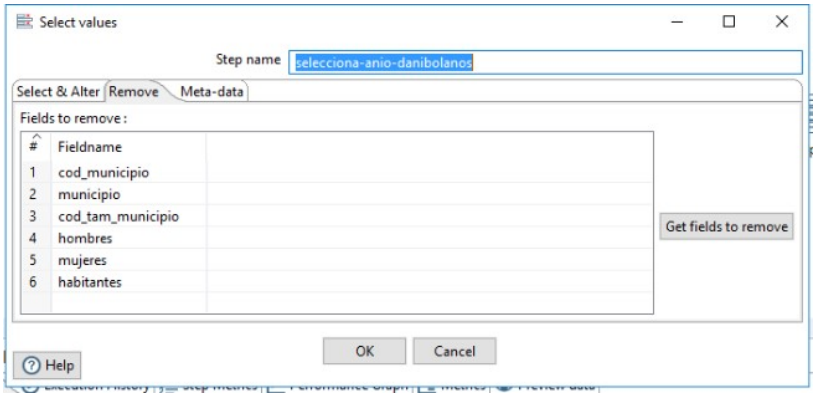

Figura 21: Eliminamos los campos innecesarios.

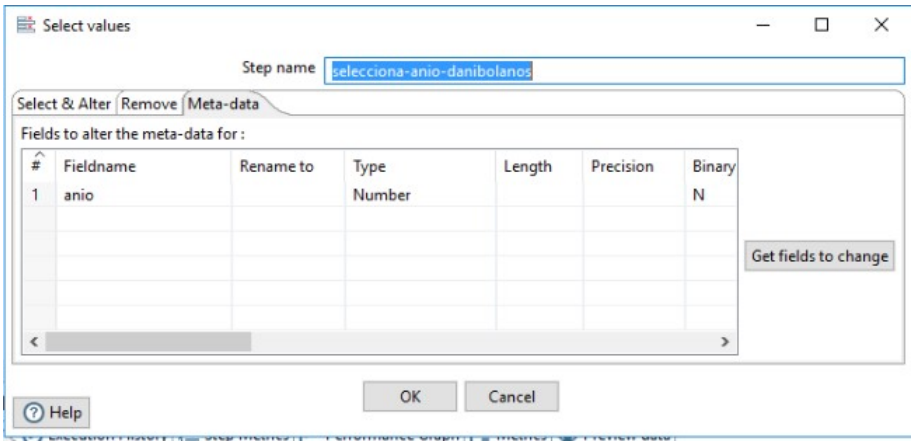

Figura 22: Cambiamos el tipo del campo **anio** a *Number*.

Cambiamos a tipo *Number* para poder hacer cálculos con la información que nos aporta para generar el campo **decada**.

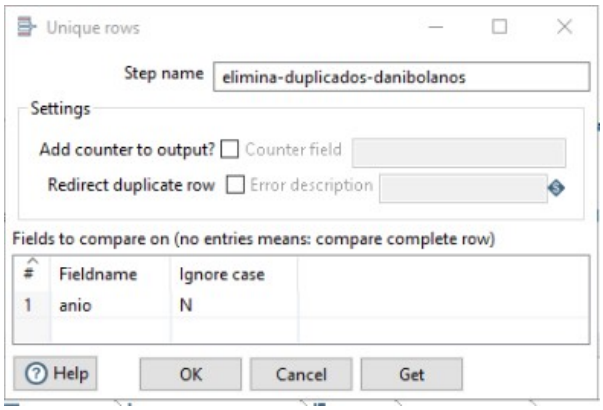

Figura 23: Eliminamos las filas duplicadas para el campo **anio**.

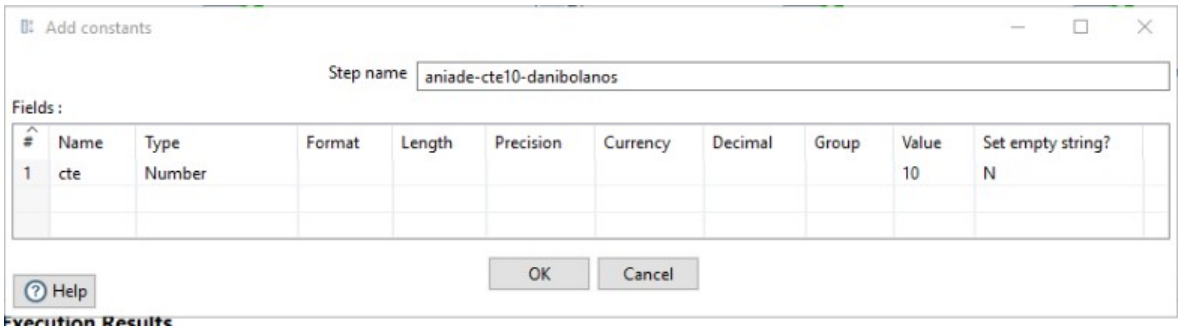

Figura 24: Creamos el campo **cte**.

Añadimos un campo **cte** y lo completamos con el valor 10 para poder usarlo posteriormente para calcular el campo **decada**.

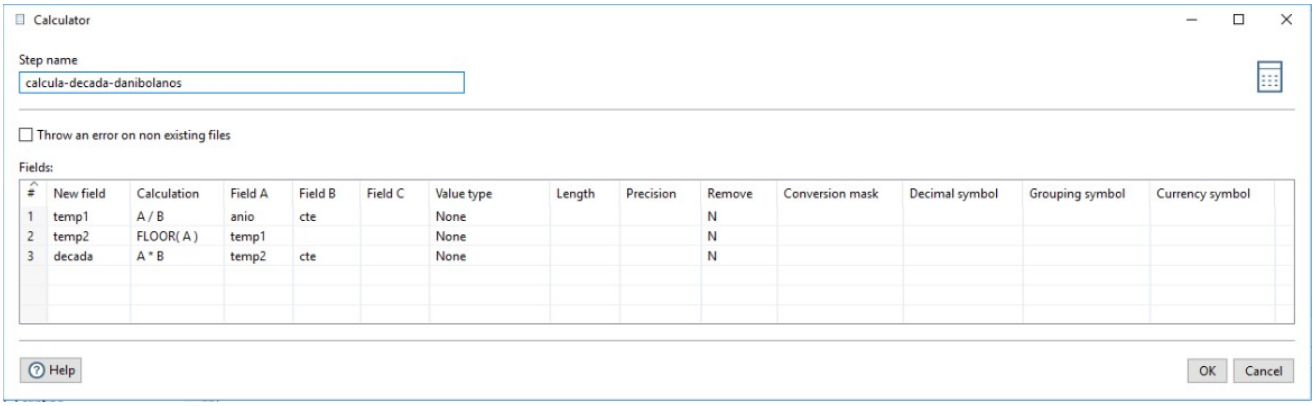

Figura 25: Calculamos el campo **decada** a partir del resto.

Calculamos el campo **decada** a partir de la fórmula: $floor(\frac{anio}{10}) \cdot 10$.

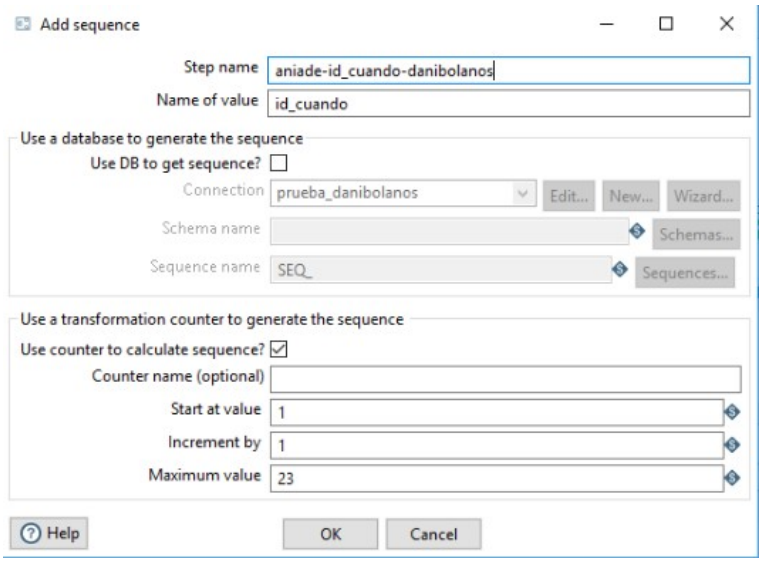

Figura 26: Añadimos el campo **id_cuando**.

Generamos los valores del campo de 1 a 23.

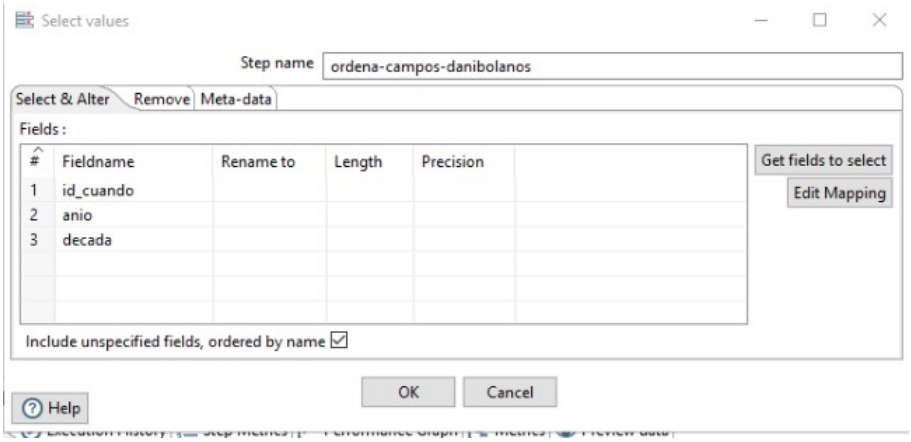

Figura 27: Reordenamos los valores para adecuarlos a la tabla **cuando**.

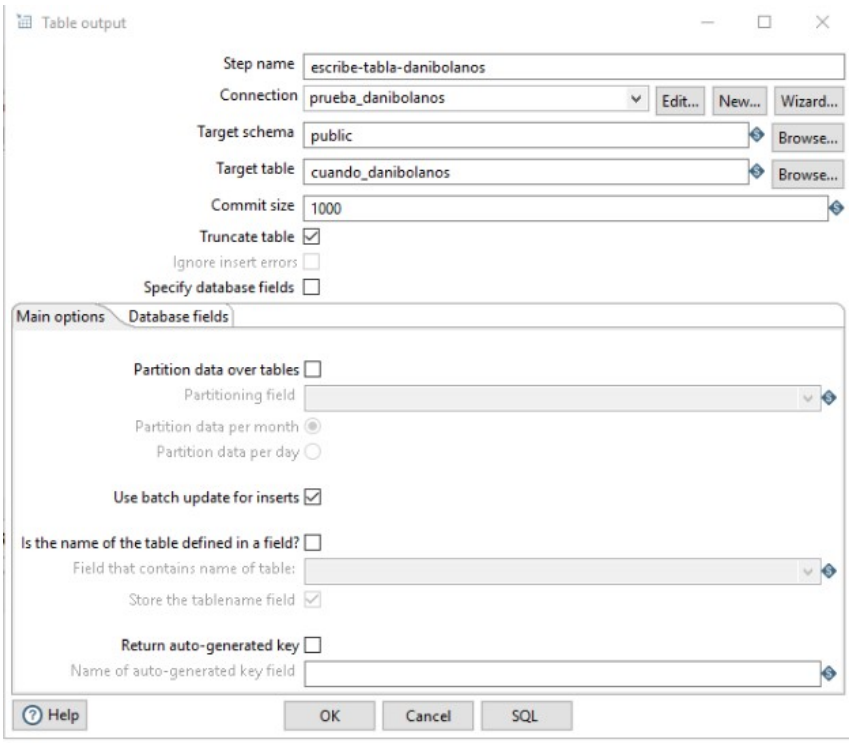

Figura 28: Escribimos los valores en la tabla.

Pulsamos el botón *SQL* para crear la tabla **cuando_danibolanos**.

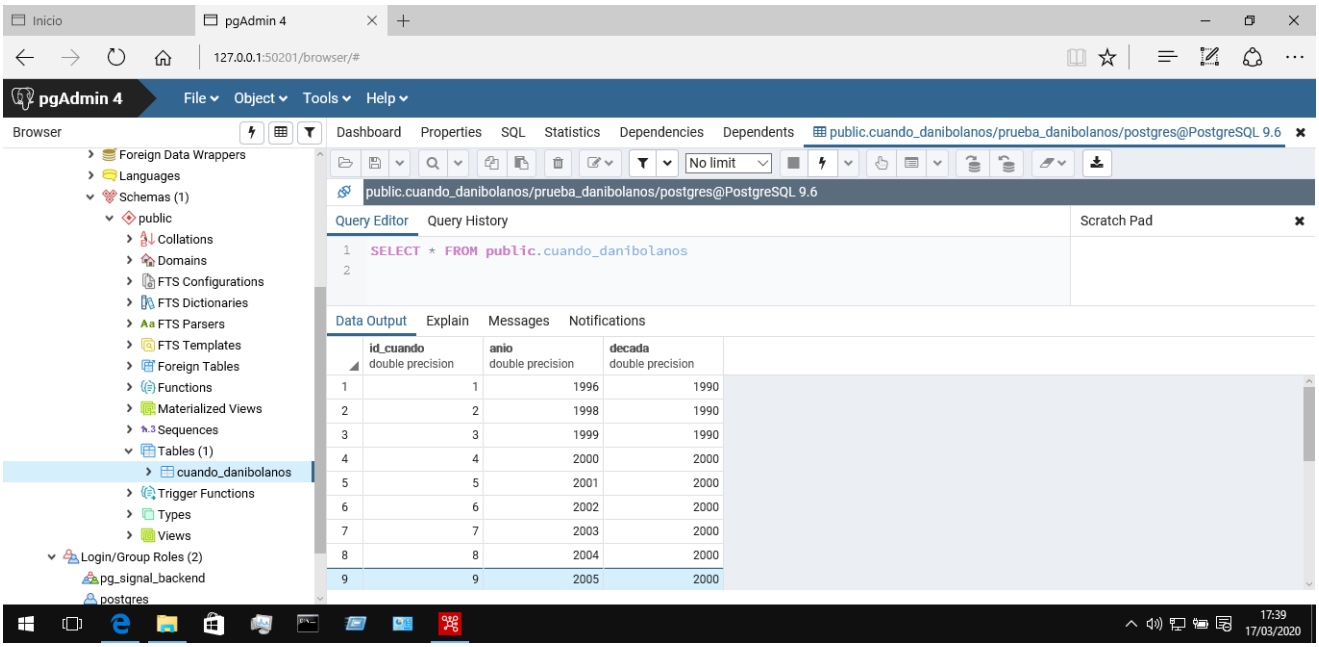

Figura 29: Vemos los resultados en la BD.

Podemos ver los resultados obtenidos para la tabla **cuando_danibolanos** y cómo se adecua a los valores del ejercicio 1.

3. Bibliografía.

Referencias

[1] Guión de prácticas de la asignatura.