PEC1



22.510 - Diseño y uso de bases de datos analíticas.

*Santiago Domínguez Collado.*

PEC 1 - Diseño y uso de bases de datos analíticas.

Parte Teoría:

**Pregunta 1:**

¿Cuál es la principal diferencia existente entre un sistema orientado al tema y uno orientado a la funcionalidad?

La principal diferencia es que los sistemas orientados a la funcionalidad, como pueden ser los almacenes operacionales, vienen precedidos por un conjunto de requerimientos, lo que hacen que se construya el sistema para satisfacer una necesidad tan concreta como conocida. Por otro lado, en los sistemas orientados a la funcionalidad, no se tiene dicha información, y es por ello por lo que se decide orientar en torno a áreas o temas.

¿A qué tipo pertenecería el Data Warehouse? Justifica tu respuesta.

El Data Warehouse es un sistema orientado al tema. En el momento del diseño de un almacén de datos, no es posible conocer las necesidades de los analistas, por lo tanto, no se puede llegar a conocer los requerimientos concretos que tiene.

**Pregunta 2:**

Enumera las diferencias estructurales existentes entre una base de datos operacional y un almacén de datos.

-Temporalidad: muy superior en el caso del almacén de datos. Los datos normalmente se almacenan entre 5 y 10 años, pero en las bases de datos operacionales, entre 1 y 2 años.

-Volumen: la temporalidad condiciona al volumen de datos, a mayor temporalidad, mayor cantidad almacenada

-Nivel de agregación: en las bases de datos operacionales suele ser único y bastante bajo. En los almacenes de datos se suelen ver distintos niveles.

-Actualización: en las bases de datos operacionales la actualización es constante, sin embargo, en los almacenes de datos se hace de manera periódica.

-Estructura: los sistemas de datos operacionales se ciñen más a una estructura relacional mientras que los almacenes de datos tienen una visión más multidimensional y dinámica.

**Pregunta 3:**

¿Cuáles son las principales diferencias entre un Data Mart y un Enterprise Data Warehouse? Enumeración y breve descripción de cada una de ellas.

-Temática: un Data Mart está diseñado para cubrir las necesidades de un departamento de la empresa, por lo tanto la temática de los datos es más específica que en el Warehouse, el cual se utiliza para almacenar toda la información de la empresa.

-Fuentes de datos: mientras que el Data Warehouse recibe información de múltiples fuentes de datos, el Data Mart la recibe de menos fuentes, en una FIC puede recibirla solo del almacén de datos.

-Tamaño: cuanta más información guardada, mayor tamaño tendrá el sistema. Los Data Warehouse pesan terabytes, mientras que los Data Marts pesa gigabytes.

-Tiempo de desarrollo: el almacén de datos es un sistema de gran complejidad que necesita años de desarrollo, mientras que los almacenes departamentales suelen llevar meses.

-Modelos de datos: los Data Marts se ciñen a un modelo relacional que garantiza las necesidades de los usuarios, mientras que los Data Warehouse multidimensional que permite una mejora en el rendimiento mediante técnicas específicas.

**Pregunta 4:**

¿Qué enfoque en la construcción de la FIC consideras más adecuado? Argumenta la respuesta.

Considero más adecuado la construcción de la FIC mediante proyectos autónomos. Desde mi perspectiva, el desarrollo mediante proyectos autónomos permite una evolución ordenada de la misma.

Si bien el planteamiento de la FIC mediante un solo proyecto puede parecer más natural o razonable, esta decisión conlleva a una complejidad raramente difícil de salvar. En el momento de construcción de la FIC no están definidos todos los requerimientos y funcionalidades por parte de los analistas, esto puede conllevar a innumerables cambios de organización, así como a múltiples retrasos en el proyecto que lo podrían llevar al fracaso fácilmente si estamos abordando la construcción como proyecto único.

**Pregunta 5:**

* El analista de datos debe ser capaz de implementar algoritmos de procesamiento de datos y definir modelos predictivos. ( V o F )

Falso. Si bien un analista muy experimentado puede realizar tareas de analíticas avanzadas, las funciones de procesado de datos y definición de modelos corresponden a los roles de Data Engineer y Data Scientist.

* Usando herramientas self-service BI varios analistas pueden obtener resultados diferentes a un mismo problema. ( V o F )

Verdadero. Dichas herramientas permiten libertad para explorar los datos y aportan un resultado más personal de los análisis, esto puede desembocar en que se obtengan distintos resultados en función de las interpretaciones que hagan los analistas.

* El proceso de eliminar espacios en blanco en un campo de texto es parte de la depuración de datos. ( V o F )

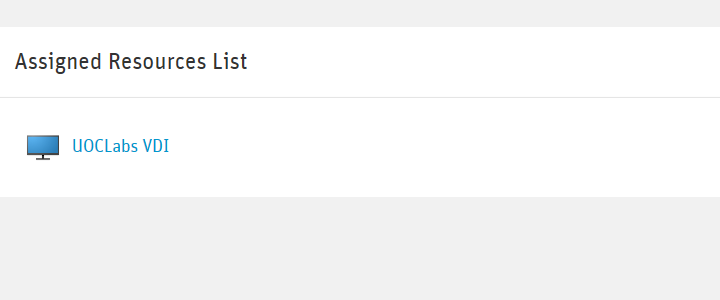
Verdadero. El espacio en blanco pude ser rellenado con un valor constante, calculado, o puede interesar dejarlo en blanco.

Parte Práctica:

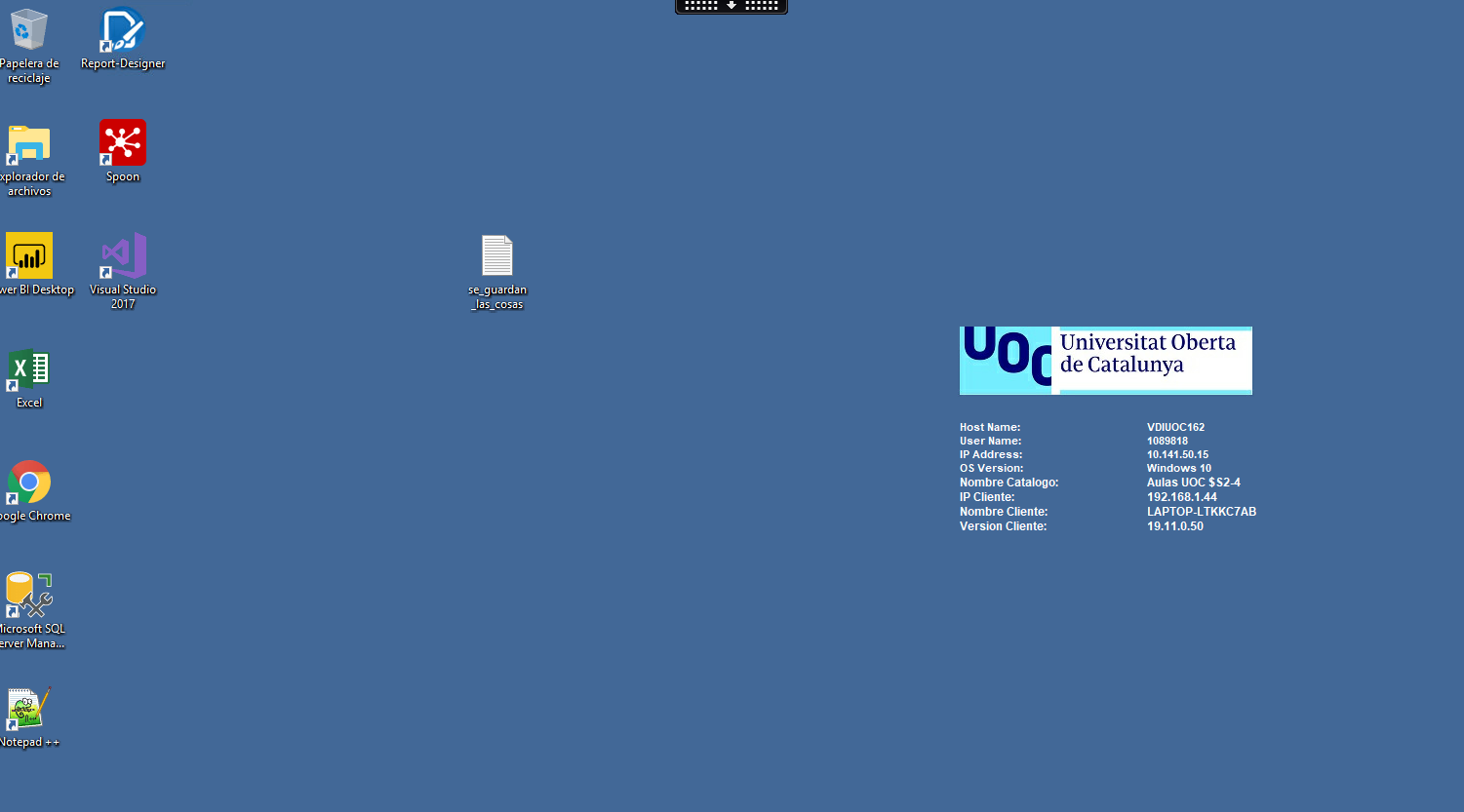
**Ejercicio 1.** Configuración del entorno de VDI.

Configurar la conexión al escritorio VDI que se utilizará durante todo el curso, tanto para realizar esta PEC (PEC1) como las tres prácticas siguientes (PRA1, PRA2, PRA3).

Tras identificarnos y descargarnos el cliente, pulsamos sobre la máquina de la asignatura.

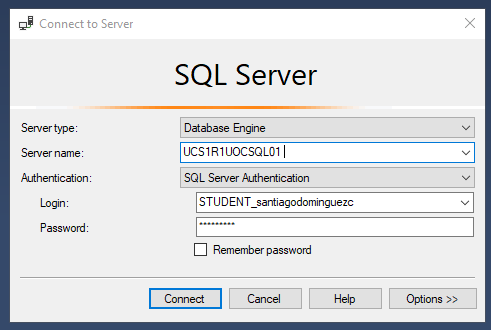


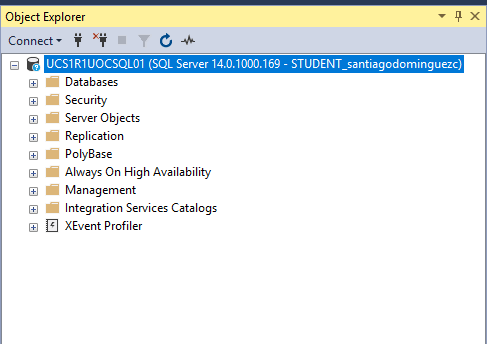
Esperamos que se inicie sesión y ya tendremos nuestro entorno con todo el software que necesitamos.



Es posible que se echen en falta algunos pasos intermedios, la razón es que yo ya accedí e hice todas las configuraciones uno de los primeros días del curso.

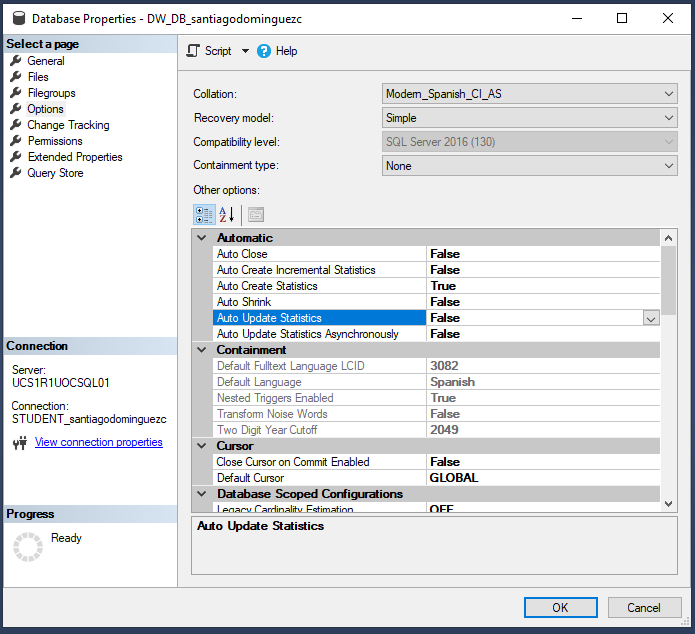
**Ejercicio 2.** Validación de la BBDD.

1. Conectar al servidor de bases de datos SQL Server.
   1. Versión de SSMS: versión Microsoft SQL Server Management Studio 17.
   2. Pantalla de conexión inicial de la consola SQL Server Management Studio. 
   3. Explorador de objetos del SSMS donde se puede ver el usuario de conexión.



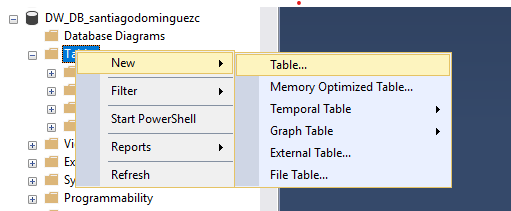
1. Configurar la base de datos DW\_DB\_XXX.

En propiedades de nuestra BD, en opciones, cambiamos “Recovery model”. También cambiamos la opción “Auto Update Stadistics” a False.

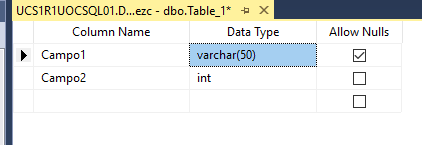


1. Crear una tabla con el nombre T\_PEC1.

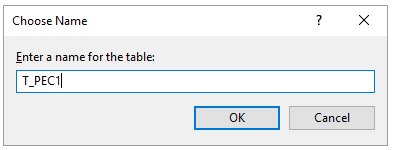
Creamos una tabla. Click derecho>New>Table



Rellenamos los campos según los requerimientos propuestos.

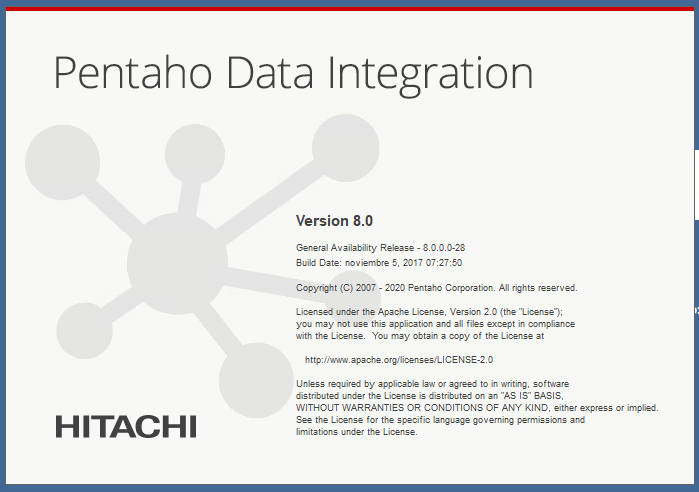


Por último, ctrl+s para guardarla como T\_PEC1.

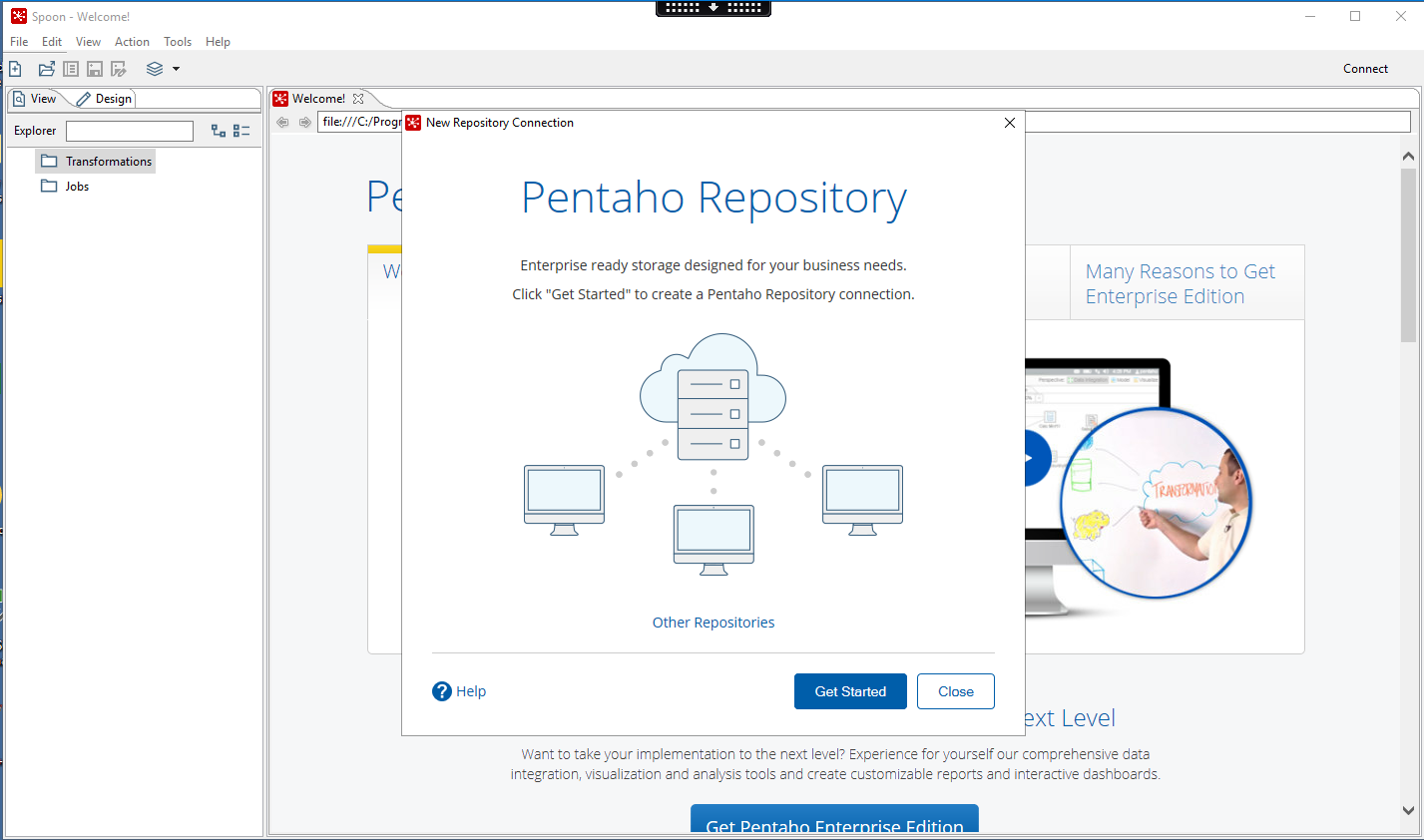


**Ejercicio 3.** Validación de la aplicación de ETL.

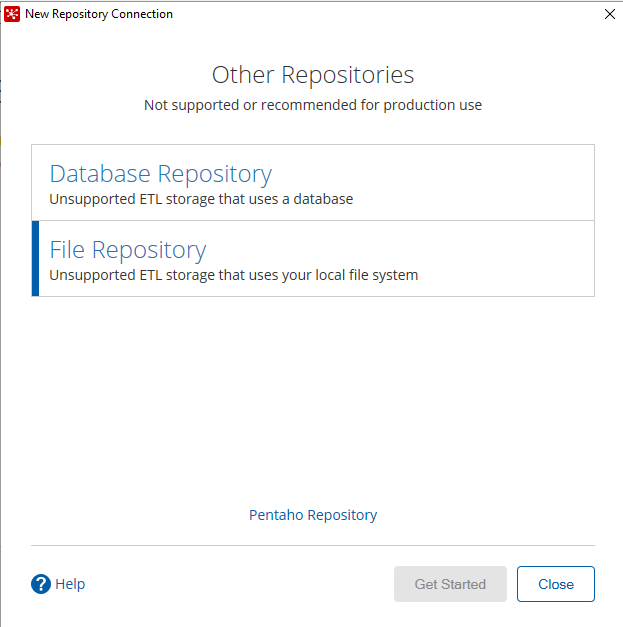
Versión del PDI. La 8.0.



Para crear un repositorio pulsamos la opción “Connect”.

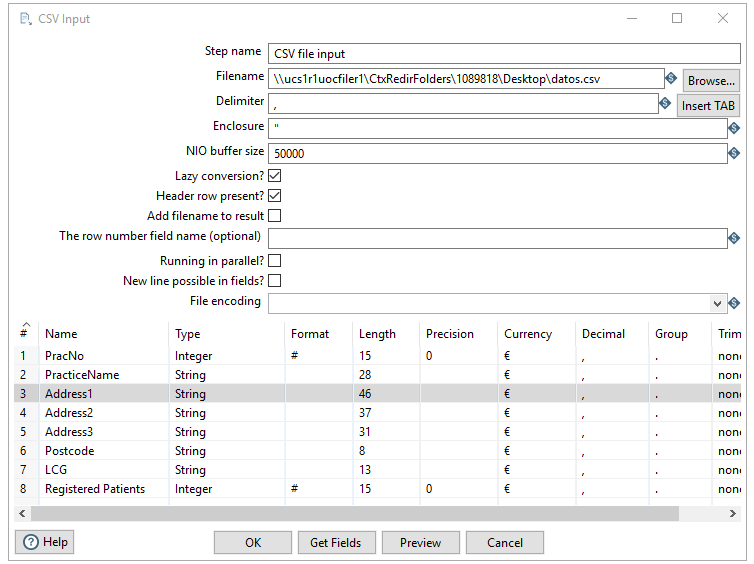


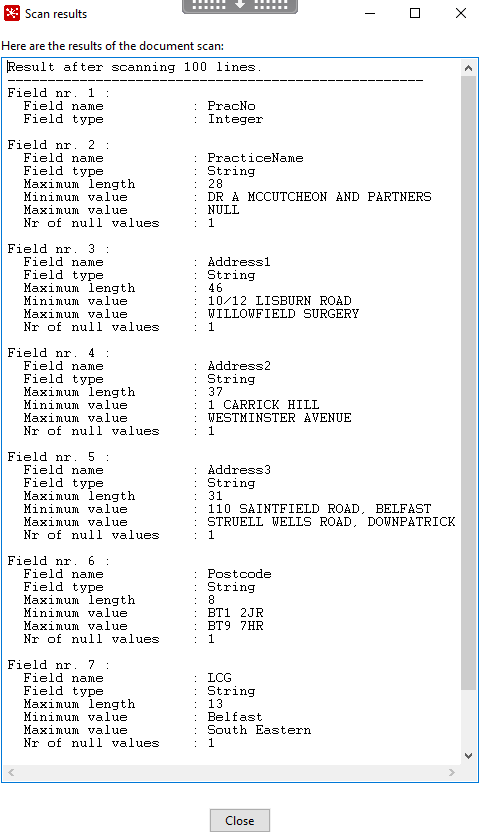
En otros repositorios, pulsamos “File Repository”.



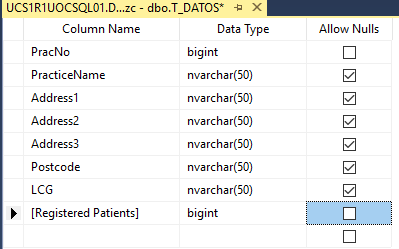
1. Análisis preliminar del origen de datos.

Para identificar todos los elementos del csv, es necesario crear una transformación de datos.csv. Pulsando la opción “get fields” tenemos la información de los campos; longitud, tipo… Salvo “PracNo” y “Registered Patients”, todos los demás campos admiten nulos.



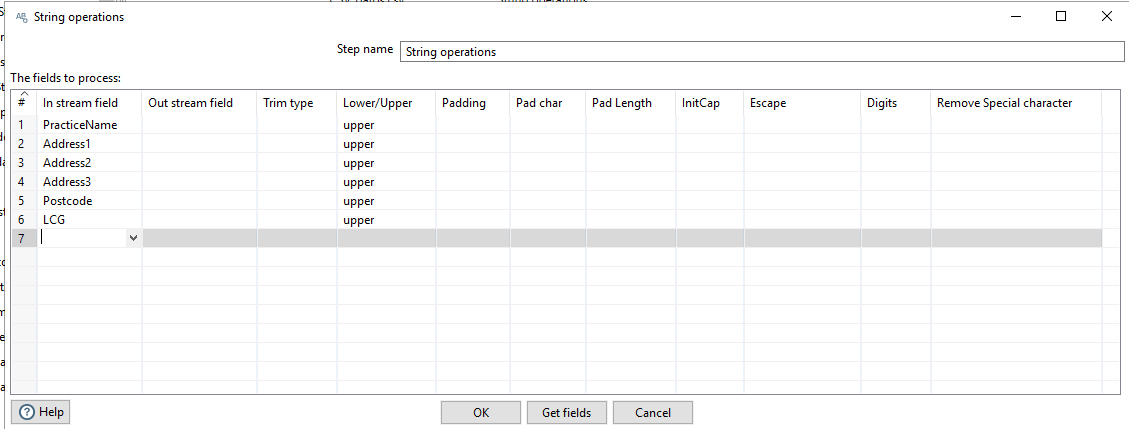


1. Crear una tabla en la BBDD con el nombre T\_DATOS.

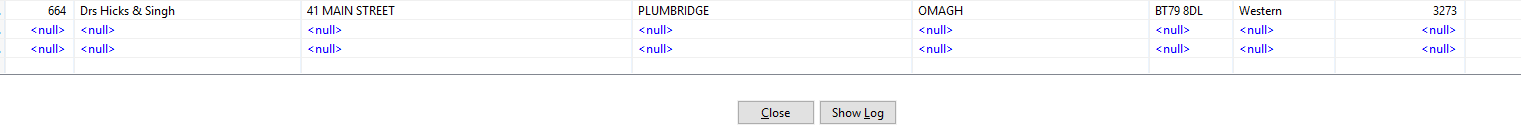


1. Crear una transformación con Spoon que realice las siguientes tareas:
2. Extraer la información del archivo CSV.
3. Transformar todos los datos de los campos de tipo texto a mayúsculas.

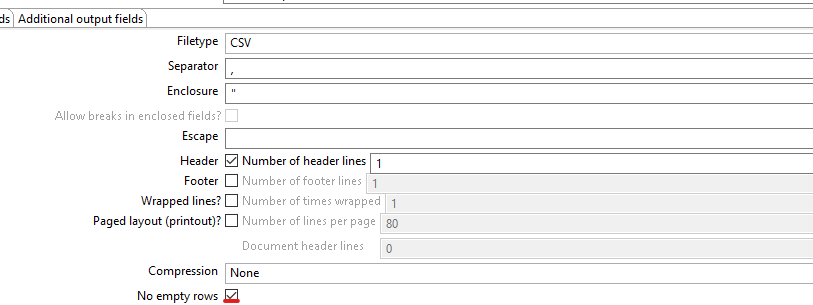
Añadimos a nuestra transformación un “String Operator” y le conectamos el step del csv. Dentro del operador, seleccionamos los campos tipo string y la opción de “upper” en todos ellos.



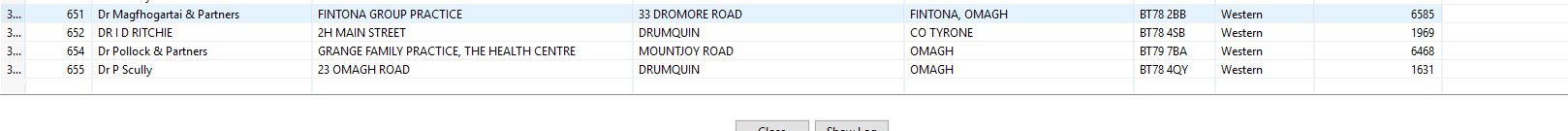
1. Corregir los errores detectados en el análisis.



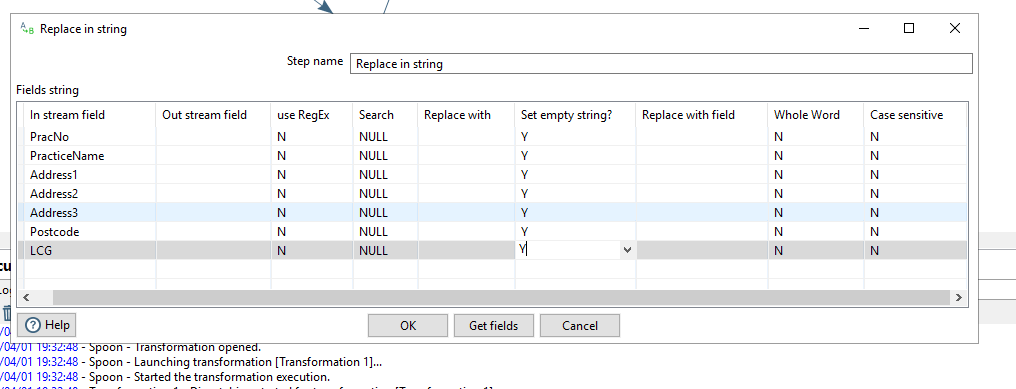
* Tras previsualizar los datos del csv, hemos observado que las dos últimas filas están vacías, procedemos a quitarlas usando la opción de “No empty rows” del step “Text File Input” en lugar del que es exclusivo para csv.



Ya no hay empty rows.

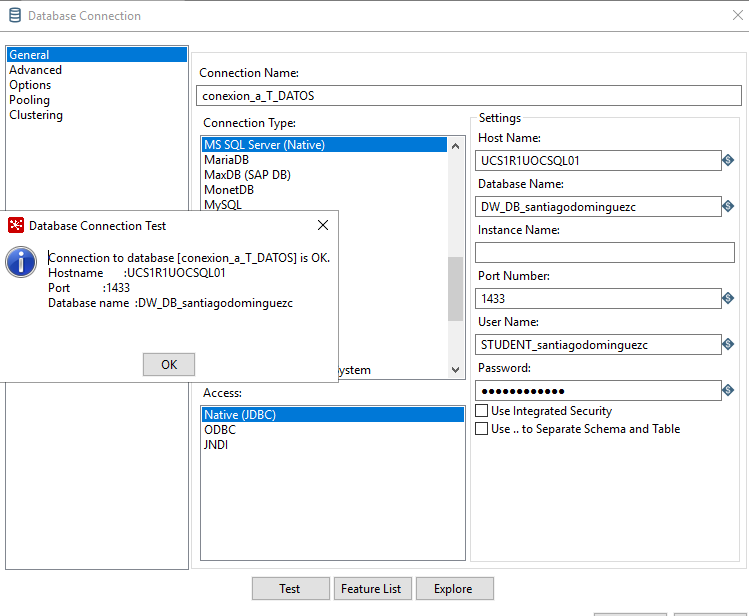


* Otro error detectado es que hay valores “null” que son string y que son propiamente de tipo null, parta ello sustituimos los primeros por string vacíos.

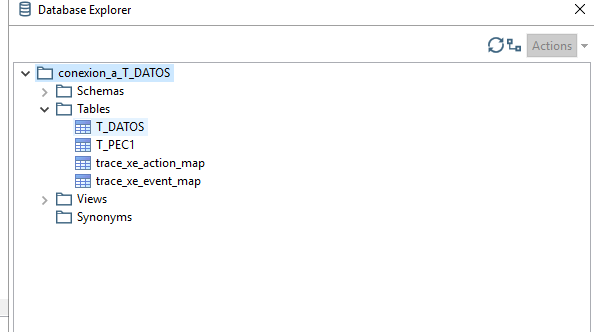


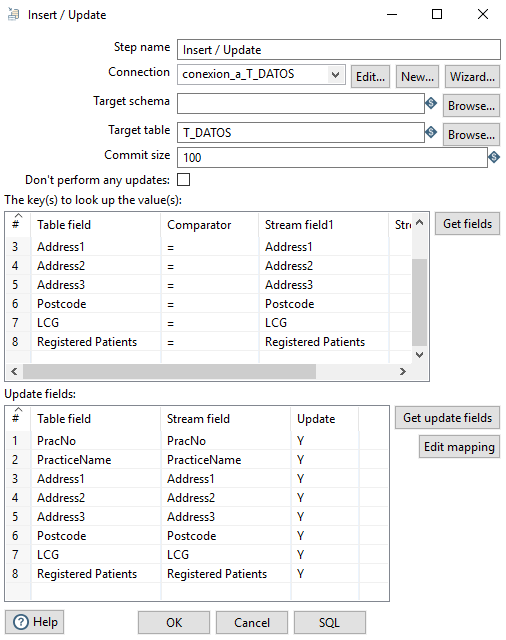
1. Cargar la información transformada en la tabla anteriormente creada.

Añadimos el elemento “Insert/Update”, configuramos una nueva conexión a nuestro MS SQL SERVER.



Tras comprobar que funciona, en la opción “Tarjet\_Table” seleccionamos nuestra tabla T\_DATOS.





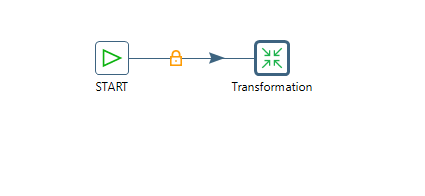
1. Realizar la carga efectiva de la tabla y validar el resultado

Al ejecutar la transformación, podemos ir a nuestra base de datos y observar que se han insertado los datos del csv y en mayúscula. Otra opción pudiera ser previsualizar los datos en una nueva transformación que reciba como input la BD.



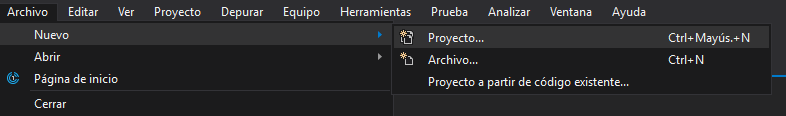
1. Crear un job que ejecute la transformación anterior.

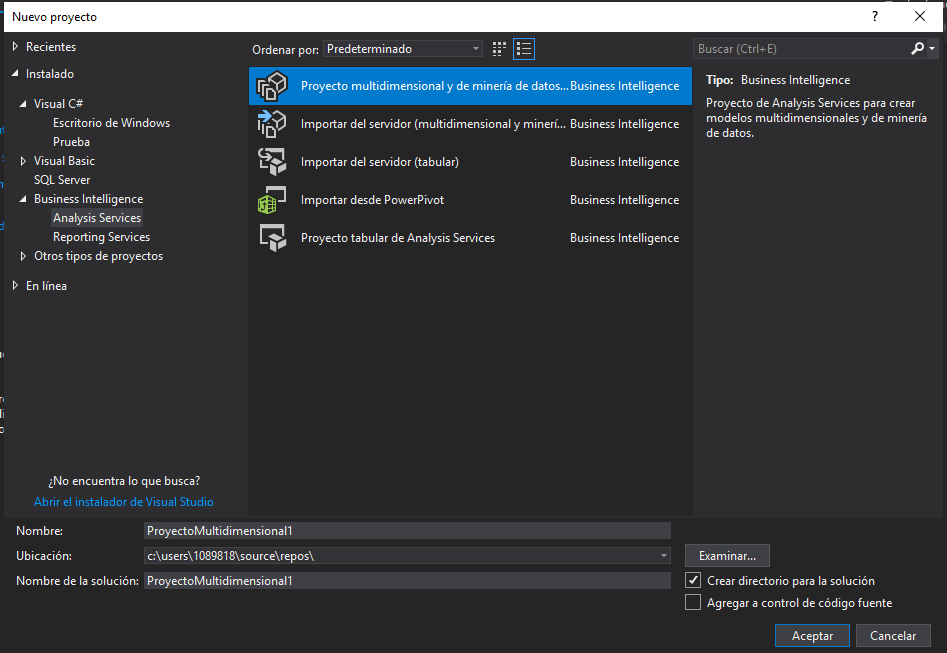
Un simple job resulta muy sencillo, lo primero es File>New>Job. Posteriormente añadimos el step “Start” y lo conectamos con el “Transformation”, en este último, introducimos el nombre de nuestra transformación. Guardamos y listo.



**Ejercicio 4.** Microsoft SQL Analysis Services.

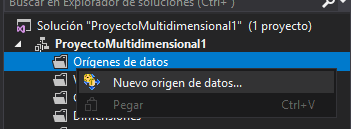
* Creamos un nuevo proyecto según las directrices del enunciado.

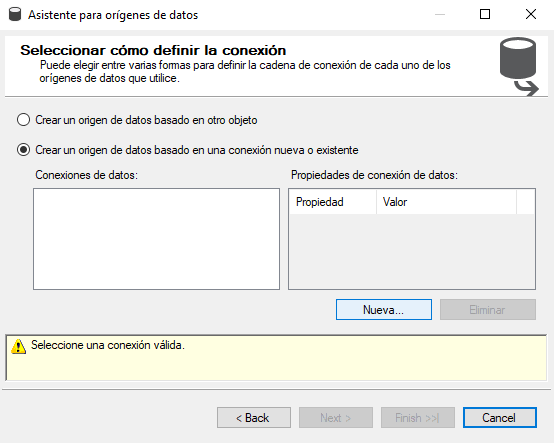




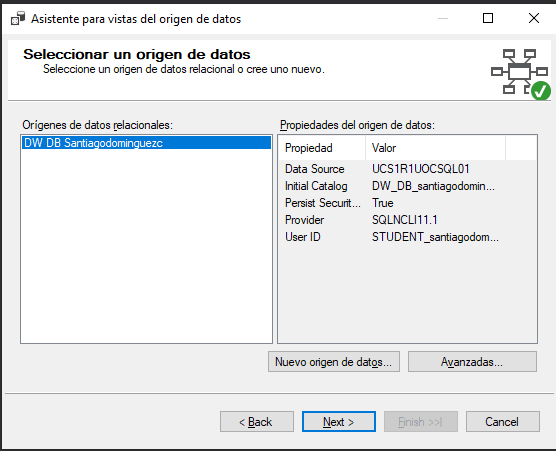
* Creamos un nuevo origen de datos.

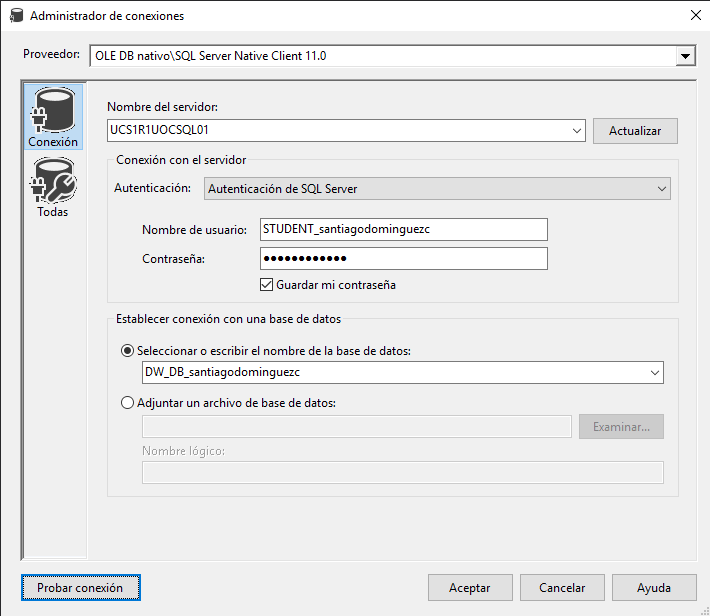
Creamos un nuevo origen de datos pulsando en el explorador de soluciones.



Seleccionamos origen basado en una conexión nueva o existente. 

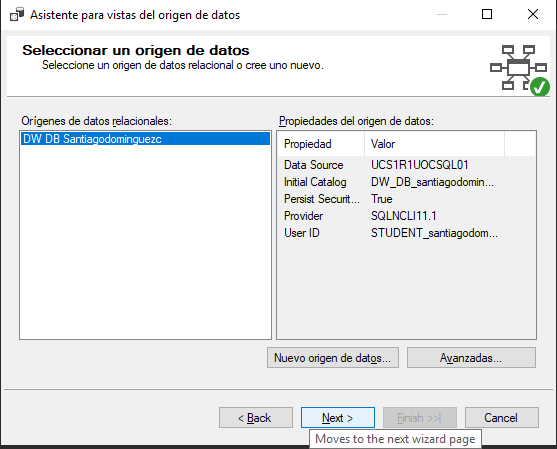
Posteriormente, introducimos los datos de nuestra BD.



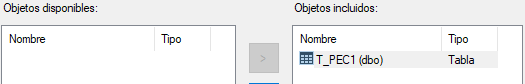


* Creación de vista de origen de datos.

De la misma forma que en la creación del origen de datos se nos abre el siguiente menú.

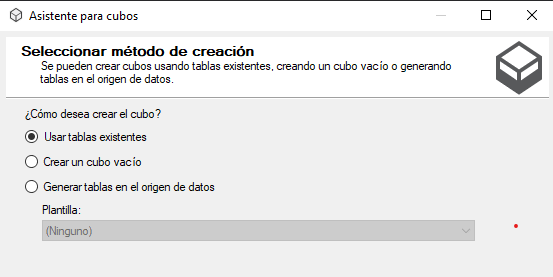


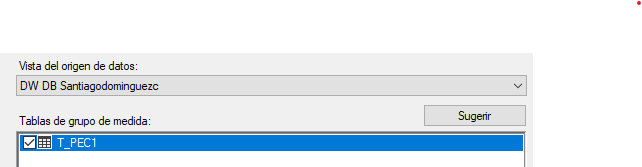
Incluimos la tabla creada y posteriormente pulsamos finalizar.



* Creación un cubo con una única tabla de hechos T\_Datos.

Creamos un cubo basado en una tabla existente.





Incluimos ambos campos, incluimos todas las opciones que nos ofrece el asistente en los siguientes pasos.



Cubo creado.

