

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 procesamiento 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 puede 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.

Assigned Resources List



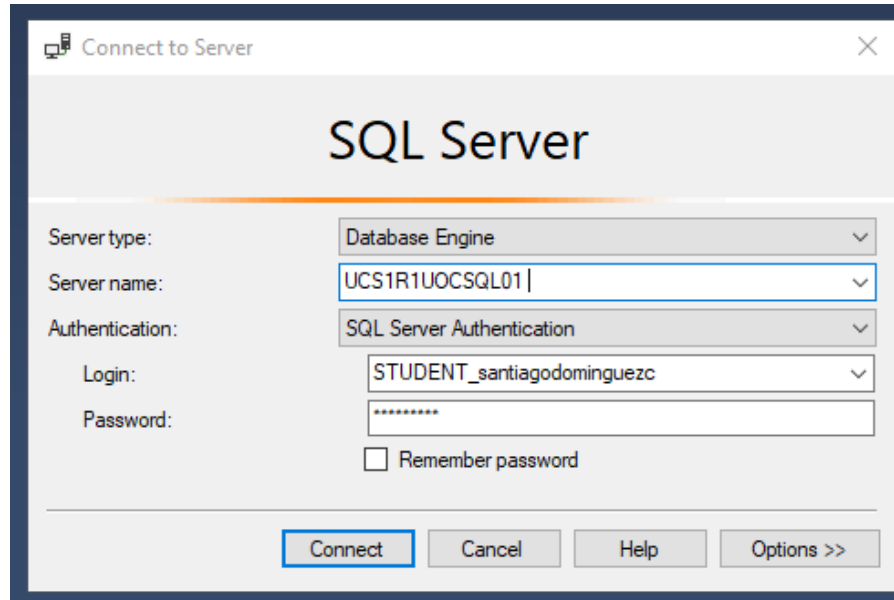
Esperamos que se inicie sesión y ya tendremos nuestro entorno con todo el software que necesitamos.



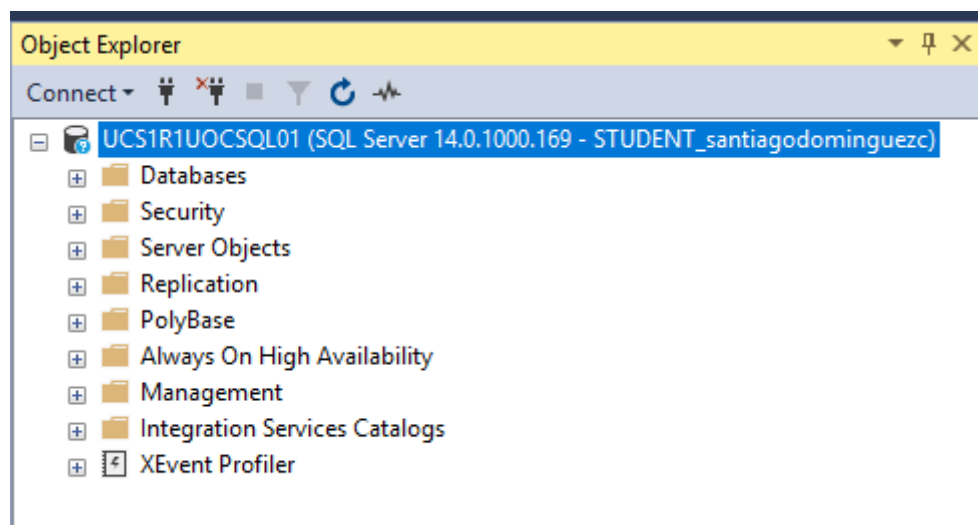
Es posible que se echen en falta algunos pasos intermedios, la razón es que yo ya accedí e hice todas las configuraciones uno de los primeros días del curso.

Ejercicio 2. Validación de la BBDD.

- 1) Conectar al servidor de bases de datos SQL Server.
 - a. Versión de SSMS: versión Microsoft SQL Server Management Studio 17.
 - b. Pantalla de conexión inicial de la consola SQL Server Management Studio.

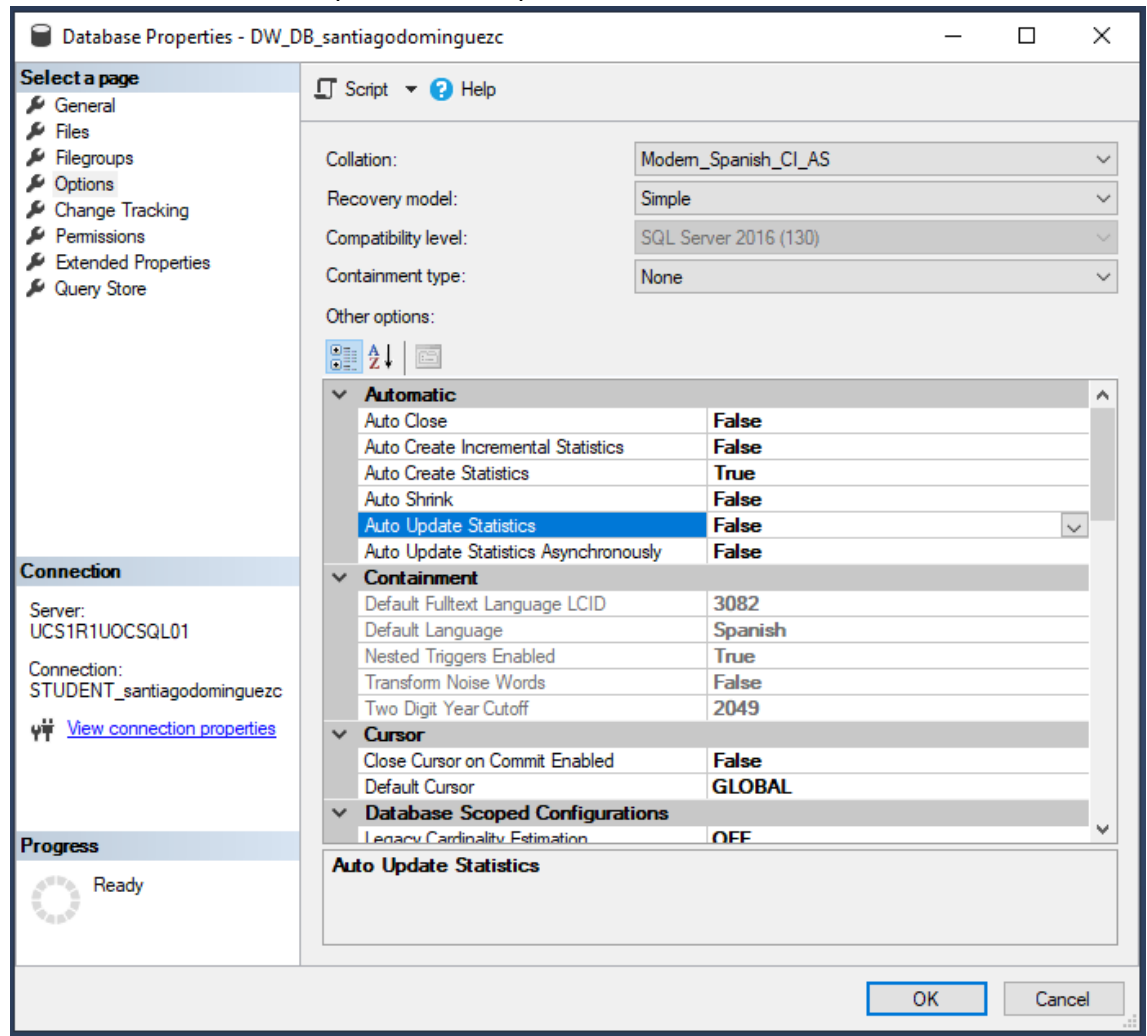


- c. Explorador de objetos del SSMS donde se puede ver el usuario de conexión.



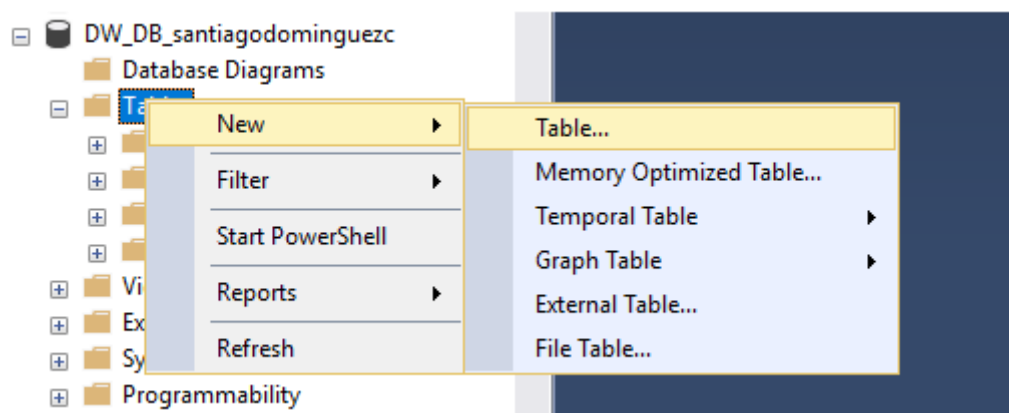
2) Configurar la base de datos DW_DB_XXX.

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



3) Crear una tabla con el nombre T_PEC1.

Creamos una tabla. Click derecho>New>Table



Rellenamos los campos según los requerimientos propuestos.

UCS1R1UOCSQL01.D...ezc - dbo.Table_1* ↵ ✕			
	Column Name	Data Type	Allow Nulls
▶	Campo1	varchar(50)	<input checked="" type="checkbox"/>
	Campo2	int	<input type="checkbox"/>
			<input type="checkbox"/>

Por último, ctrl+s para guardarla como T_PEC1.

Choose Name?×

Enter a name for the table:

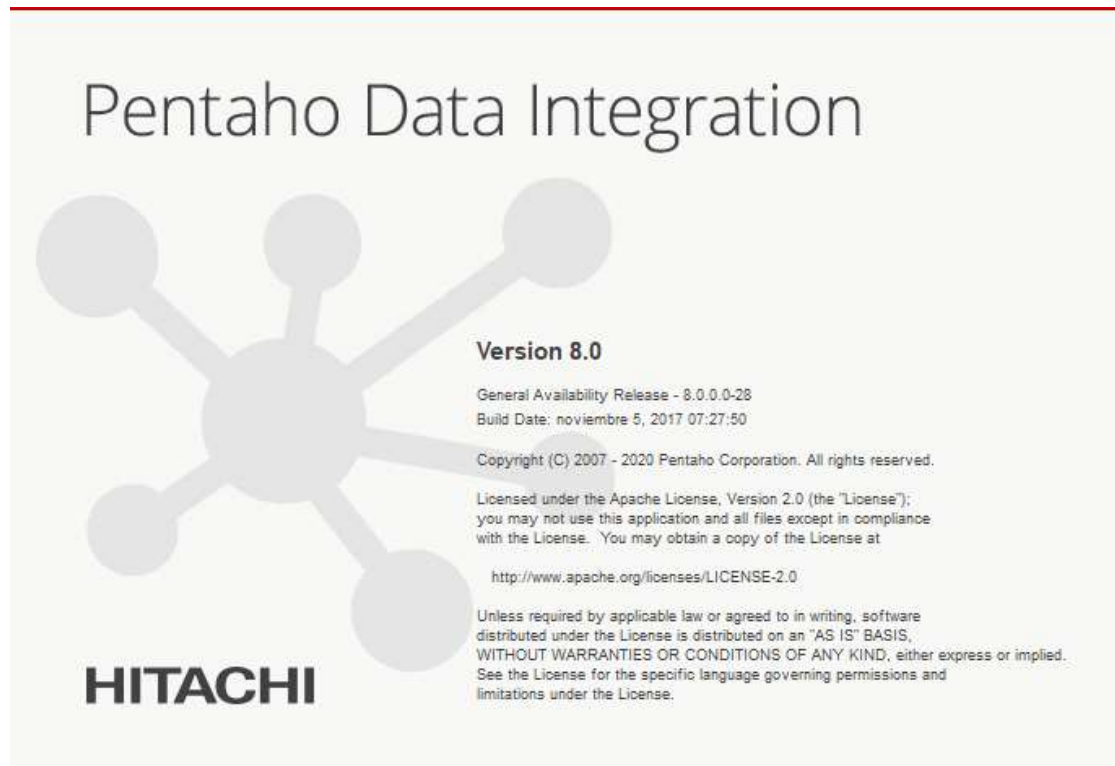
T_PEC1|

OK

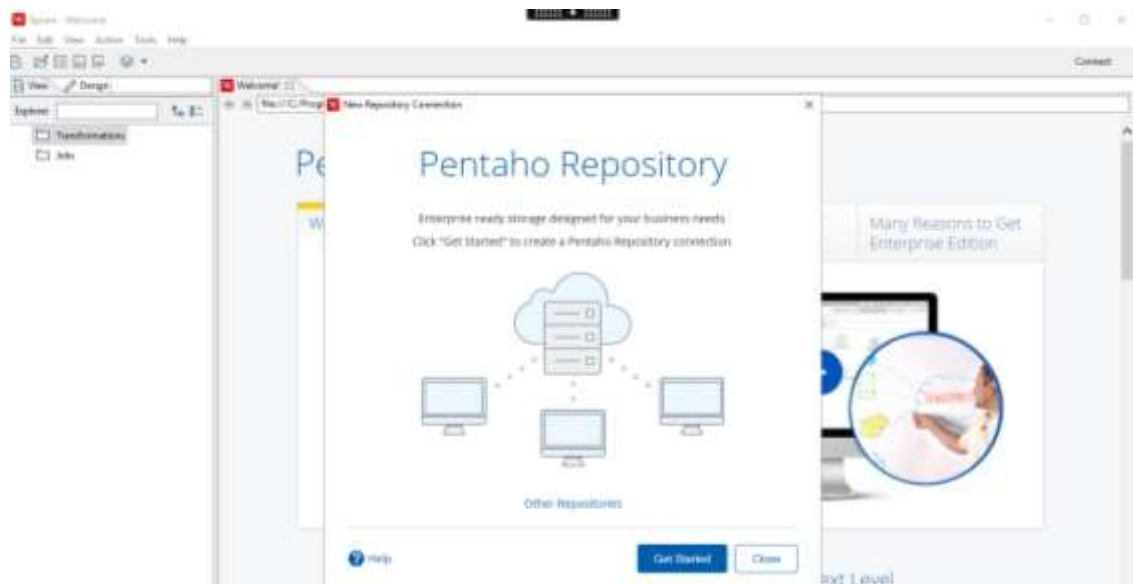
Cancel

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”.



A) 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.

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim
1	PracNo	Integer	#	15	0	€	,	.	noni
2	PracticeName	String		28		€	,	.	noni
3	Address1	String		46		€	,	.	noni
4	Address2	String		37		€	,	.	noni
5	Address3	String		31		€	,	.	noni
6	Postcode	String		8		€	,	.	noni
7	LCG	String		13		€	,	.	noni
8	Registered Patients	Integer	#	15	0	€	,	.	noni

< >

? Help OK Get Fields Preview Cancel

Scan results

Here are the results of the document scan:

Result after scanning 100 lines.

```
-----
Field nr. 1 :
  Field name      : PracNo
  Field type      : Integer

Field nr. 2 :
  Field name      : PracticeName
  Field type      : String
  Maximum length  : 28
  Minimum value   : DR A MCCUTCHEON AND PARTNERS
  Maximum value   : NULL
  Nr of null values : 1

Field nr. 3 :
  Field name      : Address1
  Field type      : String
  Maximum length  : 46
  Minimum value   : 10/12 LISBURN ROAD
  Maximum value   : WILLOWFIELD SURGERY
  Nr of null values : 1

Field nr. 4 :
  Field name      : Address2
  Field type      : String
  Maximum length  : 37
  Minimum value   : 1 CARRICK HILL
  Maximum value   : WESTMINSTER AVENUE
  Nr of null values : 1

Field nr. 5 :
  Field name      : Address3
  Field type      : String
  Maximum length  : 31
  Minimum value   : 110 SAINTFIELD ROAD, BELFAST
  Maximum value   : STRUELL WELLS ROAD, DOWNPATRICK
  Nr of null values : 1

Field nr. 6 :
  Field name      : Postcode
  Field type      : String
  Maximum length  : 8
  Minimum value   : BT1 2JR
  Maximum value   : BT9 7HR
  Nr of null values : 1

Field nr. 7 :
  Field name      : LCG
  Field type      : String
  Maximum length  : 13
  Minimum value   : Belfast
  Maximum value   : South Eastern
  Nr of null values : 1
```

Close

B) Crear una tabla en la BBDD con el nombre T_DATOS.

UCS1R1UOCSQL01.D...zc - dbo.T_DATOS*			
	Column Name	Data Type	Allow Nulls
	PracNo	bigint	<input type="checkbox"/>
	PracticeName	nvarchar(50)	<input checked="" type="checkbox"/>
	Address1	nvarchar(50)	<input checked="" type="checkbox"/>
	Address2	nvarchar(50)	<input checked="" type="checkbox"/>
	Address3	nvarchar(50)	<input checked="" type="checkbox"/>
	Postcode	nvarchar(50)	<input checked="" type="checkbox"/>
	LCG	nvarchar(50)	<input checked="" type="checkbox"/>
►	[Registered Patients]	bigint	<input type="checkbox"/>
			<input type="checkbox"/>

C) Crear una transformación con Spoon que realice las siguientes tareas:

- I. Extraer la información del archivo CSV.
- II. 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.

String operations

Step name String operations

The fields to process:

#	In stream field	Out stream field	Trim type	Lower/Upper	Padding
1	PracticeName			upper	
2	Address1			upper	
3	Address2			upper	
4	Address3			upper	
5	Postcode			upper	
6	LCG			upper	
7					

III. Corregir los errores detectados en el análisis.

004 - Drs Pineda & Singh	41 MAIN STREET	FLUORIDE	CHALK	BTR BD	Western	ECTS
→ null	→ null	→ null	→ null	→ null	→ null	→ null

Quit Show log

- 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.

ds Additional output fields

Filetype CSV

Separator ,

Enclosure "

Allow breaks in enclosed fields? ☐

Escape

Header ☒ Number of header lines 1

Footer ☐ Number of footer lines 1

Wrapped lines? ☐ Number of times wrapped 1

Paged layout (printout)? ☐ Number of lines per page 80

Document header lines 0

Compression None

No empty rows ☒

Ya no hay empty rows.

811	Dr Maghrajani & Partners	PRATYAKSH PRACTICE	10 DEBORAH ROAD	BRATONA, ONTARIO	8776 388	Western	8585
812	DR J F FETCHE	21 SAUND STREET	BRAMPTON	ON L7R5M8	8776 438	Western	1990
813	Dr Prabir & Partners	GRANGE GARDY PRACTICE THE HEALTH CENTRE	MCOWEN ROAD	CAMBRIDGE	8776 766	Western	4868
814	Dr P Scully	20 CHAMBER ROAD	BRAMPTON	ONTARIO	8776 427	Western	1887

- Otro error detectado es que hay valores "null" que son string y que son propiamente de tipo null, para ello sustituimos los primeros por string vacíos.

Replace in string

Step name Replace in string

Fields string

In stream field	Out stream field	use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole Word	Case sensitive
PracNo		N	NULL		Y		N	N
PracticeName		N	NULL		Y		N	N
Address1		N	NULL		Y		N	N
Address2		N	NULL		Y		N	N
Address3		N	NULL		Y		N	N
Postcode		N	NULL		Y		N	N
LCG		N	NULL		Y		N	N

Help OK Get Fields Cancel

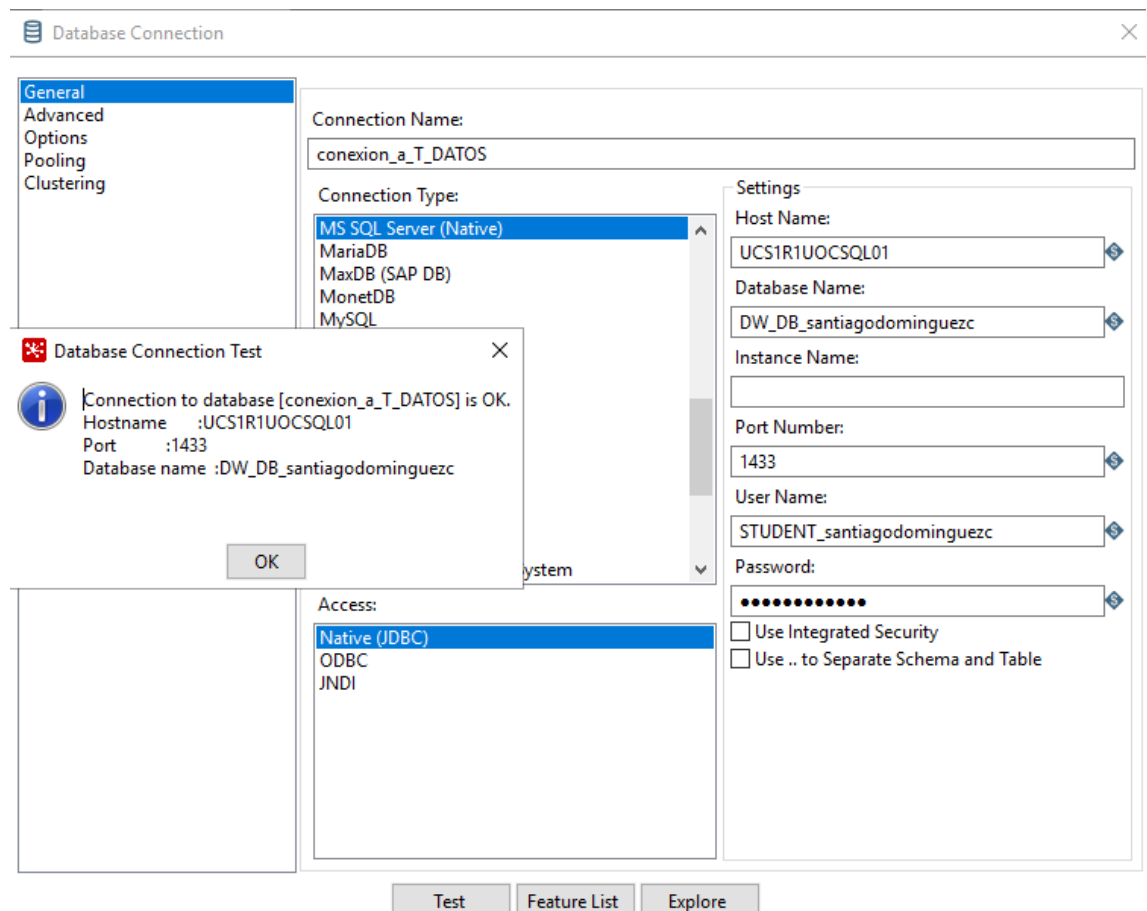
04/01 19:32:48 - Spoon - Transformation opened.

04/01 19:32:48 - Spoon - Launching transformation [Transformation 1]...

04/01 19:32:48 - Spoon - Started the transformation execution.

IV. 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.



Insert / Update

Step name

Connection

Target schema

Target table

Commit size

Don't perform any updates: ☐

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

#	Table field	Comparator	Stream field1	Str	Get fields
3	Address1	=	Address1		
4	Address2	=	Address2		
5	Address3	=	Address3		
6	Postcode	=	Postcode		
7	LCG	=	LCG		
8	Registered Patients	=	Registered Patients		

Update fields:

#	Table field	Stream field	Update	Get update fields
1	PracNo	PracNo	Y	
2	PracticeName	PracticeName	Y	
3	Address1	Address1	Y	
4	Address2	Address2	Y	
5	Address3	Address3	Y	
6	Postcode	Postcode	Y	
7	LCG	LCG	Y	
8	Registered Patients	Registered Patients	Y	

Edit mapping

V. 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.

	PracNo	PracticeName	Address1	Address2	Address3	Postcode	LOG	Regi
169	345	DR JOHNSTON & PARTNERS	BALLYMONEY FAMILY PRACTICE	BALLYMONEY HEALTH CENTRE	218 NEWVAL ROAD, B.	BT53 6...	NORTH...	122
170	346	DR SHANNON & PARTNERS	LODGE HEALTH	20 LODGE MANOR	COLERAINE	BT52 1JX	NORTH...	102
171	348	DR LYNCH & PARTNERS	THE COUNTRY MEDICAL PRACTICE	122 BALLINLEA ROAD	ARMOY, BALLYMONEY	BT53 8TY	NORTH...	606
172	350	DRS WEE & BROWN	6 PRIESTLAND ROAD	BUSHMILLS	NULL	BT57 8...	NORTH...	288
173	351	DR. JS BAILIE AND PARTNERS	PORTRUSH MEDICAL CENTRE	17 DUNLUCE AVENUE	PORTRUSH	BT56 8...	NORTH...	733
174	352	DR KERR & PARTNERS	GARVAGH HEALTH CENTRE	110 MAIN STREET	GARVAGH, COLERAINE	BT51 5AE	NORTH...	594
175	354	DRS J HARLEY & D HARLEY	THE FAMILY PRACTICE	6 LEVER ROAD	PORTSTEWART	BT55 7EF	NORTH...	318
176	355	DR MCGURK & PARTNERS	THE HEALTH CENTRE	36 GARVAGH ROAD	KILREA, COLERAINE	BT51 5...	NORTH...	679
177	356	DR TURNER & PARTNERS	4 MOUNTSANDAL ROAD	COLERAINE	CO LONDONDERRY	BT52 1JB	NORTH...	114
178	357	DR MCSARRAN & PARTNERS	GLENS OF ANTRIM MEDICAL CENTRE	2 GORTACLEE ROAD	CUSHENDALL, BALL...	BT44 0TE	NORTH...	514
179	358	DRS NUTT & SIBERRY	LIFROCK	69 SEA ROAD	CASTLEROCK, COLE...	BT51 4...	NORTH...	269
180	360	DRS HENDERSON & YOUNG	THE HEALTH CENTRE	10 MONEYLECK ROAD	RASHARKIN, BALLY...	BT44 8...	NORTH...	430
181	361	DR FANNIN & PARTNERS	THE HEALTH CENTRE	ROBINSON MEMORIAL HOSPITAL	BALLYMONEY	BT53 6...	NORTH...	533
182	366	NULL	THE HEALTH CENTRE	DALRIADA HOSPITAL	BALLYCASTLE	BT54 6BA	NORTH...	351
183	367	DRS MCCURDY & ARMSTRONG	THE PROCESS MEDICAL CENTRE	56 MAIN STREET	CLOUGHMILLS, BALL...	BT44 9LF	NORTH...	440
184	368	DRS BELL & PARTNERS	THE HEALTH CENTRE	210 BALLYMONEY	BALLYMONEY	BT53 6...	NORTH...	470

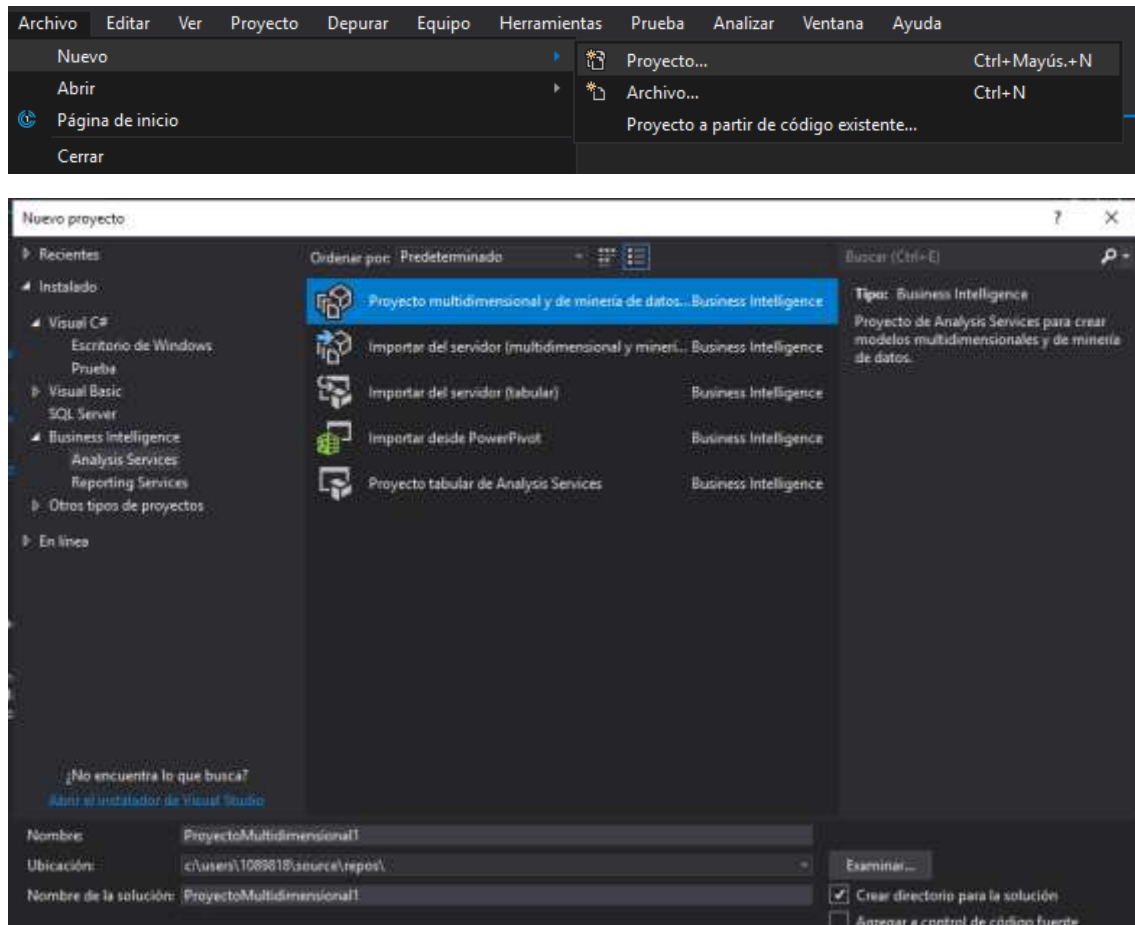
D) Crear un job que ejecute la transformación anterior.

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.



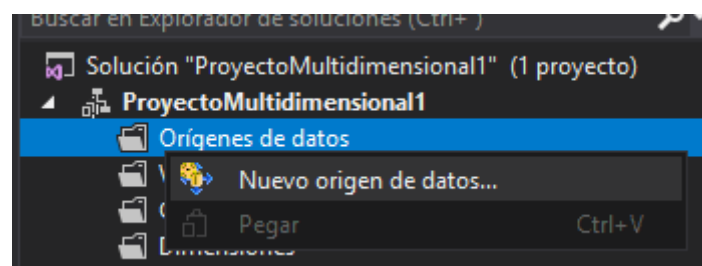
Ejercicio 4. Microsoft SQL Analysis Services.

- 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.


 Asistente para orígenes de datos

Seleccionar cómo definir la conexión
Puede elegir entre varias formas para definir la cadena de conexión de cada uno de los orígenes de datos que utilice.

☐ Crear un origen de datos basado en otro objeto

☒ Crear un origen de datos basado en una conexión nueva o existente

Posteriormente, introducimos los datos de nuestra BD.

 Asistente para vistas del origen de datos

Seleccionar un origen de datos
Seleccione un origen de datos relacional o cree uno nuevo.

Orígenes de datos relacionales:
DW DB Santiagodominguezc

Propiedades del origen de datos:

Propiedad	Valor
Data Source	UCS1R1UOCSQL01
Initial Catalog	DW_DB_santiagodomin...
Persist Securit...	True
Provider	SQLNCLI11.1
User ID	STUDENT_santiagodom...

< Back Next > Finish >> Cancel

Administrador de conexiones

Proveedor: OLE DB nativo\SQL Server Native Client 11.0

Nombre del servidor: UCS1R1UOCSQL01 Actualizar

Conexión con el servidor

Autenticación: Autenticación de SQL Server

Nombre de usuario: STUDENT_santiagodominguezc

Contraseña: ●●●●●●●●

☒ Guardar mi contraseña

Establecer conexión con una base de datos

☒ Seleccionar o escribir el nombre de la base de datos:

DW_DB_santiagodominguezc

☐ Adjuntar un archivo de base de datos:

Examinar...

Nombre lógico:

- 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.

Asistente para vistas del origen de datos

Seleccionar un origen de datos
 Seleccione un origen de datos relacional o cree uno nuevo.

Orígenes de datos relacionales:

DW DB Santiagodominguezc

Propiedades del origen de datos:

Propiedad	Valor
Data Source	UCS1R1UOCSQL01
Initial Catalog	DW_DB_santiagodomin...
Persist Securit...	True
Provider	SQLNCLI11.1
User ID	STUDENT_santiagodom...

Nuevo origen de datos... Avanzadas...

Incluimos la tabla creada y posteriormente pulsamos finalizar.

Objetos disponibles:			Objetos incluidos:	
Nombre	Tipo		Nombre	Tipo
		>	T_PEC1 (dbo)	Tabla

- Creación un cubo con una única tabla de hechos T_Datos.

Creamos un cubo basado en una tabla existente.

Asistente para cubos

Seleccionar método de creación
Se pueden crear cubos usando tablas existentes, creando un cubo vacío o generando tablas en el origen de datos.

¿Cómo desea crear el cubo?

☒ Usar tablas existentes

☐ Crear un cubo vacío

☐ Generar tablas en el origen de datos

Plantilla:
(Ninguno)

Vista del origen de datos:
DW DB Santiagodominguezc

Tablas de grupo de medida:
☒ T_PEC1

Sugerir

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

T_PEC1

☒ Campo1

☒ Campo2

Cubo creado.

