Solución: "El diseño de un almacén de datos para la gestión de la hospitalización de un hospital general básico"

Josep Curto Díaz

PID_00211704

UOC
Universitat Oberta de Catalunya

Índice

Introducción

La resolución de esta actividad, "El diseño de un almacén de datos para la gestión de la hospitalización de un hospital general básico", consta de cinco partes.

1) Estudio preliminar: se analiza la viabilidad y rentabilidad del proyecto.

2) Entorno de trabajo: se revisa el entorno de trabajo.

3) Diseño del *data warehouse*: el objetivo final es proponer un almacén de datos que permita llevar a cabo el seguimiento estratégico y operativo del área de hospitalización de un hospital general básico.

4) Carga de datos: se diseñan e implementan los procesos de extracción, transformación y carga de datos en el almacén de datos propuesto en el punto anterior.

5) Explotación de datos: se diseña un modelo MOLAP para el análisis multidimensional de la información disponible en el almacén de datos gracias al punto anterior.

1. Estudio preliminar

El primer paso en todo proyecto –y esto se aplica también a aquellos vinculados a las tecnologías de la información– consiste en determinar la viabilidad y la rentabilidad del mismo.

En nuestro caso particular, el diseño de un almacén de datos para la gestión de la hospitalización de un hospital general básico, es necesario hacer un análisis del periodo de retorno de inversión, de la tasa interna de rentabilidad, del descuento de flujos financieros y del retorno de la inversión.

El primer paso es recopilar la información que nos proporcionan en el enunciado. La información proporcionada es la siguiente:

- Se considera la evolución del proyecto durante cinco años.
- El primer año el sistema tendrá 25 usuarios. El número de usuarios se incrementará progresivamente de 10 en 10 hasta llegar a los 65 usuarios en el quinto año.
- La inversión inicial del proyecto es de 76.000 euros.
- Los costes de los siguientes años incluyen mantenimiento, evolución y nuevas licencias y se mantienen constantes, ascendiendo a 16.700 euros por año.
- Los beneficios estimados para el primer año son de 20.000 euros, 25.000 euros en el segundo y 35.000 euros en los siguientes años.
- La tasa de descuento es del 10% a lo largo de los años considerados.

Los datos proporcionados no permiten entrar en un detalle pormenorizado de las líneas que componen costes y beneficios, si bien, en un caso real, para llegar a estas cantidades se habrán determinado las principales magnitudes como, por ejemplo, el coste de hardware, software, implementación, mantenimiento o ampliación del proyecto.

El enunciado proporciona una información que puede llevar a diferentes interpretaciones y, por lo tanto, generar distintos escenarios de análisis.

Contenido complementario

Determinar correctamente las magnitudes que definen el proyecto para calcular de manera precisa su viabilidad y rentabilidad es una de las fases más relevantes del mismo, y resulta crucial para determinar de manera previa y posterior el éxito del proyecto de la implementación del almacén de datos.

1.1. Escenario 1: inversión inicial en primer año

Se considera que la inversión inicial se hace el primer año[1] del periodo (en vez de usar un año 0), y que los beneficios que proporciona el proyecto se refieren a beneficio bruto.

(1)Este enfoque supone que la inversión inicial no supone un pago por adelantado. Además, en el año 1 no se aplica la tasa de descuento.

Esta información se resume en la siguiente tabla, y nos permite calcular el flujo de caja durante los primeros cinco años del proyecto.

	Año 1	Año 2	Año 3	Año 4	Año 5
Inversión inicial	76.000 €				
Usuarios	25	35	45	55	65
Costes[1]		16.700 €	16.700 €	16.700 €	16.700 €
Beneficios estimados	20.000 €	25.000 €	35.000 €	35.000 €	35.000 €
Tasa de descuento	10%	10%	10%	10%	10%
Flujo de caja[2]	–56.000 €	8.300 €	18.300 €	18.300 €	18.300 €

1. A partir del segundo año, e incluyen mantenimiento, nuevas licencias y evolución.
2. En este ejemplo, el flujo de caja es igual al beneficio neto que resulta de la resta de los beneficios brutos estimados del proyecto menos los costes.

Por consiguiente, podemos calcular las métricas que nos permitirán tomar una decisión sobre el proyecto:

	Año 1	Año 2	Año 3	Año 4	Año 5
VAN[1]	–50.909,09 €	–41.487,60 €	–25.409,47 €	–10.792,98 €	2.494,74 €
TIR[2]		–80%	–27%	–1%	12%
ROI[3]	–74%	–63%	–39%	–15%	9%

1. Acrónimo de valor actual neto, cuyo significado es el valor presente de un determinado número de flujos de caja futuros.
2. Acrónimo de tasa interna de retorno, cuyo significado es el promedio geométrico de los rendimientos futuros esperados de dicha inversión: tasa de descuento con la que el valor actual neto o valor presente neto (VAN o VPN) es igual a cero.
3. Acrónimo de *return on investment* (retorno de la inversión)

¿Cómo hemos hecho los cálculos?

a) Para calcular el flujo de caja del proyecto en un determinado año t, que llamaremos F_t, se lleva a cabo la diferencia entre beneficios y costes (o en el caso del primer año, la inversión inicial). De este modo, el flujo de caja del primer año es:

$$F_1 = 20.000 - 76.000 = \quad -56.0000$$

b) Para calcular el VAN en el periodo *t*, es necesario utilizar la fórmula siguiente.

$$VAN_t = \sum_{i=1}^{t} \frac{F_i}{(1+TIR_i)^i}$$

donde asumimos que TIR_i es constante e igual a la tasa de descuento, y donde *Fi* es el flujo de caja.

c) De este modo, para calcular el VAN para el primer año simplemente aplicamos la fórmula:

$$VAN_t = \frac{F_1}{1+10\%} = -\frac{56.000}{1+10\%} = -50.909,09$$

d) El cálculo de la TIR es más complejo. Supone resolver la ecuación resultante de igualar el VAN del periodo t a cero.

$$VAN_t = \sum_{i=1}^{t} \frac{F_i}{(1+TIR)^i} = 0$$

La ecuación en el año 1 tiene sentido por definición. La TIR se calcula para el segundo año. Para nuestro caso particular en el año 2, la ecuación resultante es:

$$0 = \frac{F_1}{1+TIR} + \frac{F_2}{(1+TIR)^2}$$

La ecuación anterior es resoluble por métodos algebraicos, y proporciona el valor siguiente:

$$TIR = -\left(1+\frac{F_2}{F_1}\right)$$

Y si sustituimos, nos proporciona el valor –41.487,60. Como es posible deducir a medida que hay más años, la ecuación polinómica que hay que resolver resulta más compleja. Por ejemplo, para el año 3 deberíamos resolver una ecuación de tercer grado (para la que aún hay fórmula). De este modo, se usa software como Excel[2] para resolver de manera automática este tipo de cálculo usando una función.

[(2)]Las hojas de datos actuales ofrecen funciones para calcular la TIR fácilmente. Por ejemplo, con Excel la función correspondiente es TIR().

e) Para el cálculo del ROI, debemos aplicar la fórmula siguiente.

$$ROI_t = \sum_{i=1}^{t} \frac{Beneficios\ netos_i}{Inversión\ inicial}$$

Para el año 1, por lo tanto, el ROI es:

$$ROI_1 = \frac{20.000 - 76.000}{76.000} = -74\%$$

Para el año 2, el ROI es:

$$ROI_2 = \frac{-76.000 + 8.300}{76.000} = -63\%$$

f) Los valores para los años siguientes se calculan de manera equivalente.

En este caso, hemos medido el retorno de la inversión respecto al primer año, pero es posible hacer este análisis dinámico incluyendo los costes asumidos en cada año y, por lo tanto, considerar la inversión total. Como es posible imaginar, esta segunda manera de trabajar es más ajustada a la realidad. En caso de que calculáramos el ROI para que los costes de cada año se tuvieran en cuenta, los cálculos son ligeramente distintos.

a) Para el año 1, por lo tanto, el ROI es:

$$ROI_1 = \frac{20.000 - 76.000}{76.000} = -47\%$$

b) Para el año 2, el ROI es:

$$ROI_2 = \frac{(20.000 - 76.000) + 8.300}{76.000 + 16.700} = -51\%$$

Y el resultado final es el siguiente:

	Año 1	Año 2	Año 3	Año 4	Año 5
Beneficios	20.000 €	25.000 €	35.000 €	35.000 €	35.000 €
Inversión inicial	76.000 €				
Costes		16.700 €	16.700 €	16.700 €	16.700 €
Beneficio neto	–56.000 €	8.300 €	18.300 €	18.300 €	18.300 €
ROI	–74%	–51%	–27%	–9%	5%

Por consiguiente, podemos calcular las métricas que nos permitirán tomar una decisión sobre el proyecto:

- El cálculo de la TIR nos indica que solo a partir del quinto año se superará el mínimo de rentabilidad exigido por la organización (el 10%), condición necesaria para dar viabilidad al proyecto. De este modo, el periodo mínimo que necesita este proyecto es de unos cinco años[3].

(3)En el caso de tener información por meses, se podría ajustar con mayor precisión el periodo necesario para que el proyecto llegue a la rentabilidad establecida por la tasa de descuento.

- El cálculo de la TIR también nos indica un par de detalles importantes:
  - Será necesario a lo largo del proyecto identificar otros beneficios potenciales que puedan derivarse del almacén de datos y que permitan conseguir antes la rentabilidad exigida.
  - Será necesaria una gestión excelente del proyecto, pues cualquier impacto en tiempo, recursos y dinero puede incidir negativamente sobre la rentabilidad.

- El cálculo del VAN, como es posible imaginar, está vinculado con el cálculo de la TIR. Si uno nos indica el periodo en el que llegamos a la rentabilidad exigida, el otro señala la cantidad de ganancias. En nuestro caso particular, en el quinto año, la inversión producirá ganancias por encima de la rentabilidad exigida (10%) y, por lo tanto, el proyecto puede aceptarse en este año.

- Por otro lado, por las hipótesis de esta actividad, sabemos que a partir del segundo año se genera beneficio neto positivo. Sin embargo, hasta el quinto año realmente no se genera un ROI positivo para el proyecto, tanto si lo calculamos contra la inversión continua como contra la inversión total. Esto nos da una magnitud que indica que este tipo de proyectos son a largo plazo y, por este motivo, es necesario identificar áreas de alto impacto para asegurar la continuidad de los mismos a largo plazo.

Teniendo en cuenta estos resultados, los comentarios iniciales para la gerencia son que el proyecto consigue rentabilidad a los cinco años y es viable, aunque con un retorno de la inversión limitado. En este sentido, las recomendaciones más prudentes son:

- Revisar el escenario actual para identificar otros posibles beneficios, así como potenciales reducciones de costes para reducir los riesgos en la viabilidad y la rentabilidad.

- En el caso de que sea posible, conseguir otros escenarios para la implementación del proyecto (en los que cambian la solución, el hardware, el proveedor y los costes) para tener otras opciones que comparar.

- En el caso de que no sea posible tener otros escenarios, iniciar el proyecto con personal experto en la gestión de proyectos de implementación de almacenes de datos, para evitar impactos en la rentabilidad y viabilidad del proyecto.

1.2. Escenario 2: inversión inicial en el año 0

Se considera que la inversión inicial se hace en el año 0[4] y que los beneficios del proyecto se refieren a beneficio bruto.

(4)Esto significa que se requiere hacer un pago por adelantado del proyecto, pero este incluye ya los costes de mantenimiento del primer año.

Siguiendo los mismos argumentos que en el escenario 1, llegamos a la tabla siguiente.

	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
Inversión inicial	76.000 €					
Usuarios		25	35	45	55	65
Costes			16.700 €	16.700 €	16.700 €	16.700 €
Beneficios estimados		20.000 €	25.000 €	35.000 €	35.000 €	35.000 €
Tasa de descuento		10%	10%	10%	10%	10%
Flujo de caja	–76.000 €	20.000 €	8.300 €	18.300 €	18.300 €	18.300 €

Con esta información, podemos hacer el cálculo de las métricas del proyecto[5]: TIR, VAN y ROI.

(5)En este caso, haremos solo el cálculo del ROI basado en los costes acumulados de cada año, como en el segundo cálculo del escenario 1.

	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
VAN		–57.818,18 €	–50958.68 €	–37.209,62 €	–24.710,47 €	–13.347,61 €
TIR		–74%	–51%	–21%	–6%	3%
ROI		–74%	–51%	–27%	–9%	5%

1.3. Comparando los escenarios de viabilidad

Aunque el primer escenario presenta que el proyecto es viable en el quinto año superando apenas la rentabilidad demandada al proyecto (12%), el segundo escenario muestra que el proyecto puede no ser viable en el quinto año.

En este escenario, aunque se consigue generar ROI positivo, al tener en cuenta la tasa de descuento queda claro que no se alcanza la rentabilidad demandada en el proyecto (queda en un 3%, contra el 10% requerido).

¿Qué es lo que sucede? En definitiva, este segundo análisis nos indica que para que sea realmente viable, el proyecto debe impactar de manera mucho más profunda en el negocio y generar un mayor beneficio bruto (en forma de ingresos, ahorros, etc.).

Una pregunta natural, considerando los dos escenarios, es la siguiente: ¿es realmente viable el proyecto? Teniendo en cuenta los dos análisis y sus resultados, debemos comentar que el proyecto tiene asociado un riesgo importante para la organización si hay alguna desviación en tiempo, recursos u otros costes. De manera natural, el responsable TI conjuntamente con el responsable financiero podrían asumir el segundo escenario como el más ajustado a la realidad y, por lo tanto, declinar el proyecto, a no ser que se pueda asegurar un mayor impacto en la organización.

2. Entorno de trabajo

Para el presente caso, se ha proporcionado una imagen virtual alojada en Amazon Web Services (AWS), con las características siguientes.

- Memoria RAM: 2 GB
- Sistema operativo: Windows 7
- Espacio: 8 GB
- Base de datos: Oracle XE 11g
- Herramienta ETL: Microsoft SQL Server Integration Services 2012
- Herramienta MOLAP: Microsoft SQL Server Analysis Services 2012
- Herramienta de diseño de ETL y Analysis Services: Visual Studio 2012

El procedimiento de acceso, uso y configuración queda fuera del alcance de este documento, ya que forma parte de la asignatura a la que está vinculada esta actividad. Cuando sea necesario, se explicarán detalles que mejoren o complementen el desarrollo de la actividad, siempre y cuando la faciliten.

3. Diseño del *data warehouse*

Tal y como se indica en el enunciado de la actividad, el diseño de la factoría de información en esta actividad se centrará solo en el área de hospitalización, para facilitar la creación de una solución de principio a fin.

3.1. Análisis de requerimientos

El análisis de requerimientos se basa en identificar las necesidades específicas que tiene una particular organización respecto al análisis de la información. La información recopilada en esta fase permite definir *a posteriori* los procesos de negocio que hay que analizar y las perspectivas de negocio respecto a las que interesa analizar el rendimiento de la organización.

Normalmente, en esta fase se debe ser previsor y pensar más allá de las necesidades actuales para cubrir las futuras. Por ejemplo, si los usuarios han mostrado interés por un análisis mensual pero hay información de más detalle, es necesario identificar si la solución de la factoría de información puede beneficiarse a futuro de información más desglosada o es suficiente con ciertos niveles agregados de información.

Es necesario comentar que, aunque para esta actividad el análisis de requisitos se fundamenta en el enunciado proporcionado y se complementa con las fuentes de datos, esta fase se debe hacer entrevistando al cliente y analizando las necesidades de información de la organización que se cubren en la actualidad y las que debe cubrir el proyecto a futuro. Por este motivo, como consultores del proyecto es conveniente hacer un diagnóstico y una conceptualización adecuada del proyecto. Es decir, identificar claramente los objetivos y el contenido del mismo.

En este caso, el objetivo es diseñar un almacén de datos que permita mejorar la gestión del área de hospitalización. En el caso del contenido, el proyecto incluye la creación e implementación de un modelo relacional, el diseño e implementación de procesos ETL, el diseño e implementación del modelo OLAP y, por último, el diseño de las consultas mínimas establecidas en el enunciado.

En un proyecto real, es necesario tener en cuenta criterios de negocio y gestión de proyectos para, por un lado, generar valor en la organización y, por otro, lograr un proyecto de éxito.

En nuestro caso particular del área de hospitalización de un hospital general básico, es de primordial interés para el centro gestionar de manera eficiente el comportamiento y el uso de esta área.

Esto se traduce en que, dentro de las necesidades de información por parte de la dirección clínica y el propio centro, podemos identificar las siguientes.

1) Conocer la evolución del uso del área de hospitalización, que consiste en conocer el uso que han hecho los ciudadanos de los servicios de hospitalización a lo largo de tiempo: cuántos pacientes ha tenido el área, cuál ha sido su episodio, qué tipo de demanda quirúrgica hay asociada a la hospitalización, cuál ha sido la evolución del paciente en caso de reingreso, etc.

2) Esta evolución se debe poder analizar desde diferentes puntos de vista. Para cada paciente, en el momento de alta en el área de hospitalización se debe registrar la siguiente información, que nos proporciona diferentes perspectivas de negocio para el análisis:

- Fecha de entrada del paciente
- Servicio que prescribe la inclusión en el RDQA[6]
- Facultativo
- Prioridad clínica del paciente
- Diagnóstico de inclusión
- Procedimiento quirúrgico
- Situación del paciente
- Motivo de salida
- Fecha de salida

[6]Acrónimo de Registro de Demanda Quirúrgica.

Si se tiene en cuenta toda esta información, el sistema resultante podrá responder a múltiples preguntas que puedan generar la dirección clínica y el propio centro.

De manera específica, se pide que el sistema, como mínimo, ha de ser capaz de dar respuesta a las preguntas siguientes.

- Evolución de las hospitalizaciones
- Evolución de las hospitalizaciones por tipo de alta
- Evolución de las hospitalizaciones por meses y años
- Evolución de las hospitalizaciones por servicio del hospital

En siguientes fases de esta actividad, será necesario definir otros aspectos relevantes para este tipo de proyectos como, por ejemplo, la periodicidad de los datos (para determinar la necesidad de carga y las ventanas de carga existentes).

En este sentido, en los proyectos de diseño de factoría de información corporativa encontramos una fase inicial del proyecto en la que se lleva a cabo una carga inicial, pero *a posteriori* hay cargas incrementales que reducir y es preciso optimizar el tiempo de carga necesario.

3.2. Análisis de fuentes de datos

En los proyectos de almacenes de datos, es muy relevante analizar las fuentes de datos. Del análisis de las mismas puede desprenderse información clave para el éxito y la evolución de proyecto, así como la identificación de riesgos vinculados. Por ejemplo:

- Identificar si hay un gobierno de datos y su nivel dentro de la organización. En el caso de que no haya una política de gestión del dato, este análisis nos debe permitir identificar como mínimo la calidad del dato y los procesos en los que participa, así como otros aspectos vinculantes como la propiedad o la periodicidad.

- Descubrir, por medio de la estructura de las fuentes, que hay registros desconocidos por los usuarios de negocio que pueden aportar valor para comprender el rendimiento del área, y que deben ser comunicados a los usuarios para su evaluación.

Puesto que en nuestro caso particular no tenemos acceso a las partes interesadas del proyecto, utilizaremos el sentido común para el análisis de fuentes de datos.

Para esta actividad, contamos con solo una fuente de datos que es un fichero Excel denominado BI_UOC_Datos_Hospitalizacionv1.

Los objetivos de analizar este fichero son:

- Identificar los campos de datos contenidos.
- Identificar los campos de datos no relevantes y que se pueden eliminar.
- Identificar el tipo de campo y si son opcionales, por lo que aceptan resultados nulos.

Este fichero se ha proporcionado de manera conjunta con el enunciado de la actividad. Los principales aspectos que hay que destacar de esta fuente de información son los siguientes.

- El fichero tiene tres hojas: datos, descripción de tabla y ficha.

- La hoja de datos contiene los registros de la actividad del área de hospitalización de ejemplo.

- La hoja de descripción de tabla contiene la descripción del significado de cada una de las columnas de los registros.

- La hoja de ficha proporciona información sobre el conjunto de datos de esta actividad, en particular el periodo de tiempo que comprenden: de enero del 2012 a enero del 2013.

- La fuente de datos contiene 14.326 registros para analizar. Cada registro está formado por 40 campos, si bien no todos son obligatorios, hecho que viene determinado por el tipo de paciente y su historial.

- La información de varios de los atributos está codificada de manera numérica.

- Parte de los campos está en catalán, por lo que debemos hacer una correspondencia al español.

- Los campos que incluye la fuente de datos son los siguientes (y es posible encontrar esta información en la hoja de descripción de tabla):

Nombre atributo (en el fichero)	Nombre atributo (en castellano)	Descripción
Any alta	Año de alta	Año alta
Classe admissió	Tipo de admisión	Tipo admisión: 1 (urgente), 2 (programado)
Classe admissió hospitalització posterior	Tipo de admisión hospitalización anterior	Igual que el anterior, si se ha producido
Classe alta	Tipo de alta	Tipo alta: podéis ver tabla Altas
Data ingrés	Fecha de ingreso	Fecha ingreso
Data alta	Fecha alta	Fecha alta
Data episodi urg seg	Fecha episodio urgencia posterior	Si ha habido urgencia posterior
Data hospitalització anterior	Fecha hospitalización anterior	Solo tiene sentido si es un reingreso
Data hospitalització posterior	Fecha hospitalización posterior	Si ha habido urgencia posterior
Episodi	Episodio	Número episodio asistencia: identifica de manera unívoca

Nombre atributo (en el fichero)	Nombre atributo (en castellano)	Descripción
És acumulat any Actual alta	Es acumulado año Actual alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
És acumulat any Anterior alta	Es acumulado año Anterior alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
És any anterior alta	Es año anterior alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
És mes actual alta	Es mes actual alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
És mes actual any anterior alta	Es mes actual año anterior alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
És mes corrent any actual alta	Es mes corriente año actual alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
És mes corrent any anterior alta	Es mes corriente año anterior alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
És setmana actual alta	Es semana actual alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
És setmana any anterior alta	Es semana año anterior alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
És setmana en curs alta	Es semana en curso alta	Valores auxiliares para identificar periodo análisis: para efectuar comparativas anuales, mensuales, semanales, etc.
Grd codi	Código GRD	Código GRD
GRD descripció	Descripción GRD	Descripción código GRD
Nhc	NHC	Numero historia clínica pacientes (están modificados)
Període alta	Periodo alta	Periodo alta (mes/año). Dato auxiliar
Servei hospitalització	Servicio hospitalización	Código servicio hospitalización al alta
Servei descriptiu	Descripción servicio	Descripción código
Sexe	Sexo	0- Hombre; 1-Mujer

Nombre atributo (en el fichero)	Nombre atributo (en castellano)	Descripción
Tipus episodi	Tipo de episodio	H: hospitalización; CM: cirugía mayor ambulatoria; DO: domiciliaria
Dies hospitalització	Días hospitalización	Días totales de hospitalización
Estada en Rea_PQ (h)	Estancia en Rea_PQ (h)	Estancia en servicios especiales (en milisegundos)
Estada en reanimació	Estancia en reanimación	Estancia en servicios especiales (en milisegundos)
Estada en UCI	Estancia en UCI	Estancia en servicios especiales (en milisegundos)
Estada en UCI CCA	Estancia en UCI CCA	Estancia en servicios especiales (en milisegundos)
Estada en unitat coronaria	Estancia en unidad coronaria	Estancia en servicios especiales (en milisegundos)
Número unitats especials ha passat	Número de unidades especiales pasadas	Servicios especiales por los que ha pasado
Peso GRD	Peso GRD	Complejidad clínica, peso del GRD de cara a comparar diferentes hospitalizaciones
Temps hospitalització (h)	Tiempo hospitalización	Tiempo hospitalización (en horas)
Diagnòstic P (codi)	Diagnóstico (código)	Diagnóstico al alta
Diagnòstic P (desc)	Diagnóstico (descripción)	Descripción código alta

Además, tenemos la tabla de altas que contiene información sobre los tipos de alta:

Tipo alta	Tipo alta descripción (en fichero)	Tipo alta descripción (en castellano)	Descripción larga
1	A domicili	A domicilio	
2	Aguts/psiquiatric	Agudos/psiquiátrico	
33	ICO	ICO	Instituto Catalán de Oncología: centros de especialidad oncología
4	Socisanitari	Sociosanitario	
41	Resid. social	Resid. social	
5	Voluntària	Voluntaria	Alta voluntaria, por decisión del paciente o representante legal, sin prescripción facultativa
6	*Exitus*	*Exitus*	Fallecimiento del paciente

Tipo alta	Tipo alta descripción (en fichero)	Tipo alta descripción (en castellano)	Descripción larga
7	Fugida	Huida	Huida del paciente, se marcha sin documentación del ingreso
8	H. domicilaria	H. domicilaria	Hospitalización domiciliaria
9	N/D	N/D	

Una vez sabemos qué información contiene nuestra fuente de datos, es necesario discernir para cada campo qué tipo de dato es y cuántos niveles de información contiene. Los niveles de información definen cuánta información jerarquizada incluye un campo, por lo que permitirán a futuro definir las jerarquías de cada dimensión. Por otro lado, las fuentes de origen de datos de una organización contienen información que no tiene por qué reflejarse en el almacén de datos. La información no relevante es distinta en función de la fuente de origen. Por ejemplo, los sistemas de información como ERP o CRM contienen atributos en la base de datos que solo tienen valor para la gestión de esta aplicación, pero no tienen valor de negocio, por lo que deben omitirse. Por tanto, en nuestro caso particular debemos revisar el significado de cada campo del fichero de datos facilitado para determinar su relevancia respecto al sistema que hay que desarrollar.

El resultado del análisis de los campos anteriores se incluye en la siguiente tabla, que presenta el resultado del análisis efectuado en cada uno de los campos en términos de tipo de datos y nivel de información contenida. Esta última columna o bien presenta el nivel de información, o bien refleja si debe omitir el campo.

Nombre atributo	Tipo de dato	Nivel de información
Año de alta	Fecha	1 nivel (año)
Tipo de admisión	Número entero	1 nivel (equivalente a urgente/programada)
Tipo de admisión hospitalización anterior	Número entero	1 nivel (equivalente a urgente/programada / N/D)
Tipo de alta	Número entero	1 nivel (equivalente a 10 tipos)
Fecha de ingreso	Fecha	6 niveles (año, mes, día, hora, minuto, segundo)
Fecha alta	Fecha	6 niveles (año, mes, día, hora, minuto, segundo)
Fecha episodio urgencia posterior	Fecha	3 niveles (año, mes, día), opcional
Fecha hospitalización anterior	Fecha	3 niveles (año, mes, día), opcional
Fecha hospitalización posterior	Fecha	3 niveles (año, mes, día)

Nombre atributo	Tipo de dato	Nivel de información
Episodio	Número entero	Identificador único del episodio con 14 caracteres
Es acumulado año actual alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Es acumulado año anterior alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Es año anterior alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Es mes actual alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Es mes actual año anterior alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Es mes corriente año actual alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Es mes corriente año anterior alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Es semana actual alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Es semana año anterior alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Es semana en curso alta	Número entero	Campo para omitir, dado que el sistema se encargará de estos cálculos
Código GRD	Número entero	Identificador código GRD
Descripción GRD	Cadena	Descripción del código GRD (1 nivel)
NHC	Número entero	Numero historia clínica pacientes (enmascarado siguiendo la LOPD)
Periodo alta	Fecha	Año - mes (contenido en fecha alta), se puede omitir
Servicio hospitalización	Cadena	Identificador servicio
Descripción servicio	Cadena	Descripción servicio (1 nivel)
Sexo	Número entero	1 nivel (equivalente a hombre/mujer)
Tipo de episodio	Cadena	1 nivel (equivalente a tres tipos)
Días hospitalización	Número entero	Métrica
Estancia en Rea_PQ (h)	Número entero	Métrica (en milisegundos), opcional
Estancia en reanimación	Número entero	Métrica (en milisegundos), opcional
Estancia en UCI	Número entero	Métrica (en milisegundos), opcional
Estancia en UCI CCA	Número entero	Métrica (en milisegundos), opcional
Estancia en unidad coronaria	Número entero	Métrica (en milisegundos), opcional

Nombre atributo	Tipo de dato	Nivel de información
Número de unidades especiales pasadas	Número entero	Métrica (en enteros), opcional
Peso GRD	Número decimal	Métrica (complejidad)
Tiempo hospitalización	Número decimal	Métrica (horas)
Diagnóstico (código)	Número decimal	Código
Diagnóstico (descripción)	Cadena	Descripción diagnóstico (1 nivel)

Este análisis nos permite identificar que de los 39 campos que tiene cada registro, hay 12 que se pueden omitir. Estos campos son valores auxiliares que permiten usar el fichero Excel para crear consultas de negocio comparativas. El **modelo MOLAP** que crearemos en posteriores fases de esta actividad permite recrear este tipo de análisis comparativos.

De los 27 campos restantes, hay ocho que pueden estar vacíos o ser opcionales, puesto que para algunas hospitalizaciones el paciente no habrá pasado por todas las unidades y, por lo tanto, no tendrá asociada una estadía. En este sentido, en el momento de carga de datos podemos completar el campo con el valor nulo correspondiente.

Nombre atributo	Tipo de dato	Opcional
Año de alta	Fecha	No
Tipo de admisión	Número entero	No
Tipo de admisión hospitalización anterior	Número entero	No
Tipo de alta	Número entero	No
Fecha de ingreso	Fecha	No
Fecha alta	Fecha	No
Fecha episodio urgencia posterior	Fecha	Sí
Fecha hospitalización anterior	Fecha	Sí
Fecha hospitalización posterior	Fecha	No
Episodio	Número entero	Sí
Código GRD	Número entero	No
Descripción GRD	Cadena	No
NHC	Número entero	No
Servicio hospitalización	Cadena	No
Descripción servicio	Cadena	No
Sexo	Número entero	No

Nombre atributo	Tipo de dato	Opcional
Tipo de episodio	Cadena	No
Días hospitalización	Número entero	No
Estancia en Rea_PQ (h)	Número entero	Sí
Estancia en reanimación	Número entero	Sí
Estancia en UCI	Número entero	Sí
Estancia en UCI CCA	Número entero	Sí
Estancia en unidad coronaria	Número entero	Sí
Número de unidades especiales pasadas	Número entero	Sí
Peso GRD	Número decimal	No
Tiempo hospitalización	Número decimal	No
Diagnóstico (código)	Número decimal	No
Diagnóstico (descripción)	Cadena	No

3.3. Análisis funcional

En el momento de considerar los requisitos funcionales, es necesario tener en cuenta que cada requisito tendrá una prioridad asociada y podrá ser exigible o deseable.

En el contexto de esta actividad, los requerimientos exigibles son aquellos que demanda el enunciado, y los deseables son los que complementan la actividad.

Por otro lado, en términos de la escala de prioridades asignamos una prioridad de 1 a 4, siendo 1 completamente prioritario para la actividad y 4, no prioritario para la actividad.

A continuación, se describen los requerimientos funcionales para el diseño de una factoría de información para el área de hospitalización, bajo las consideraciones del enunciado siguiente.

Número	Requerimiento	Prioridad	Exigible / deseable
1	Se extraerá de manera adecuada la información de las fuentes de datos (considerando solo la información relevante).	1	E
2	Se creará un almacén de datos que cubra el área de hospitalización.	1	E
3	Se cargará la información en el almacén de datos.	1	E

Número	Requerimiento	Prioridad	Exigible / deseable
4	Se creará un modelo OLAP para consultas automáticas de los usuarios.	2	E
5	Se crearán un mínimo de cuatro vistas OLAP para el análisis de la hospitalización.	2	E
6	Se crearán otras dos vistas OLAP analizando las hospitalizaciones por servicio y diagnóstico.	3	D
7	Se modificará el almacén de datos para incluir las áreas de urgencias y la lista de espera.	4	D

Cabe comentar que en un caso genérico, podemos encontrar otros requerimientos funcionales como:

- Creación de procesos de calidad de datos.
- Creación de *datamarts* (si se analizan otras áreas).
- Automatizar cada proceso de carga de *datamarts* (en función de sus necesidades).
- Creación de procesos de cargas totales e incrementales.
- Se creará un soporte a los metadatos de gestión del almacén de datos, así como de los procesos ETL.

Asimismo, dado que estos sistemas frecuentemente forman parte de la implementación de un sistema de inteligencia de negocio, la lista de requerimientos funcionales sería mucho mayor.

En términos de la arquitectura funcional, tenemos los elementos siguientes.

- Las fuentes de datos están compuestas por un único fichero Excel.

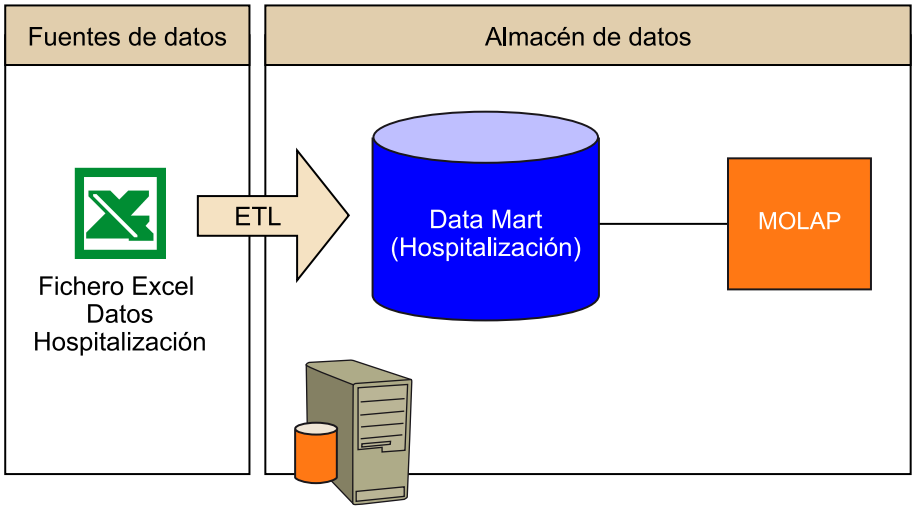
(7)Estructura de carga intermedia.

- La arquitectura de la factoría de información puede estar formada por varios elementos que estarán alojados en la misma máquina.
  - *Staging area*[7] (opcional): en el caso de tener múltiples ficheros, sería conveniente consolidarlos en una estructura de carga intermedia. En nuestro caso particular, sería creada en la misma base de datos. En todo caso, teniendo en cuenta el modelo simplificado del que partimos, se omite este paso.

– *Datamart* hospitalización: al centrarnos en una única área, es más correcto considerar que se está creando un *datamart* en lugar de un almacén de datos corporativo.

– MOLAP: a partir de la información en el *datamart* de hospitalización, se creará cubo multidimensional.

El siguiente gráfico resume los elementos de la arquitectura para esta actividad.



3.4. Diseño del modelo conceptual, lógico y físico

1) Diseño conceptual

A partir del análisis de requerimientos y del análisis de fuentes de datos, se ha identificado una tabla de hecho con 2 atributos, 9 métricas y 12 dimensiones.

La tabla de hecho corresponde al proceso de hospitalización.

Tabla de hecho	Descripción
h_hosp	Incluye el proceso de hospitalización de un paciente.

La tabla de hecho tiene los atributos y métricas siguientes.

- Atributos
  - Episodio: identificación episodio
  - NHC: número de historia clínica

- Métricas
  - d_hosp: días hospitalizado
  - est_ReaPQ: estancia en servicios especiales (en milisegundos)
  - est_sep: estancia en servicios especiales (en milisegundos)
  - est_UCI: estancia en servicios especiales (en milisegundos)

- est_UCI_CCA: estancia en servicios especiales (en milisegundos)
- est_uni_coro: estancia en servicios especiales (en milisegundos)
- num_unid: número de unidades por las que ha pasado un paciente.
- p_GRD: peso GDR que permite medir la complejidad.
- t_hosp: tiempo de hospitalización

Las dimensiones corresponden a las perspectivas de negocio sobre las que queremos analizar el proceso de hospitalización:

Dimensiones	Descripción
d_f_ingreso	Fecha de ingreso
d_f_alta	Fecha de alta
d_t_adm	Tipo de admisión
d_t_adm_posterior	Tipo de admisión posterior
d_t_alta	Tipo de alta
d_f_urg_post	Fecha de urgencia posterior
d_f_hosp_ant	Fecha de hospitalización anterior
d_f_hosp_post	Fecha de hospitalización posterior
d_GRD	GRD del episodio
d_servicio	Servicio de hospitalización en el alta
d_diagnostico	Diagnóstico en el alta
d_sexo	Sexo del paciente
d_tipo_ep	Tipo de episodio

2) Diseño lógico

Para cada dimensión se han determinado sus atributos y para las tablas de hechos, las principales métricas. De nuevo, se debe tener en cuenta la consideración anterior que detalla las dimensiones de cada tabla de hecho.

Tabla de hecho	Métricas	Atributos
h_hosp	d_hosp, est_ReaPQ, est_sep, est_UCI, est_UCI_CCA, est_uni_coro, num_unid, p_GRD, t_hosp	Episodio NHC

Las dimensiones anteriores tienen los atributos siguientes.

Dimensiones	Atributos
d_f_ingreso	Año, mes, semana, día, hora, minuto, segundo
d_f_alta	Año, mes, semana, día, hora, minuto, segundo

Dimensiones	Atributos
d_t_adm	Descripción
d_t_adm_posterior	Descripción
d_t_alta	Descripción
d_f_urg_post	Año, mes, día, semana, hora, minuto, segundo
d_f_hosp_ant	Año, mes, día, semana, hora, minuto, segundo
d_f_hosp_post	Año, mes, día, semana, hora, minuto, segundo
d_GRD	Descripción
d_servicio	Descripción
d_diagnostico	Descripción
d_sexo	Descripción
d_tipo_ep	Descripción

3) Diseño físico

En el momento de hacer el diseño físico, debemos tener en cuenta distintos aspectos.

- Tipo de base de datos con el que trabajamos, puesto que cada una de las mismas tiene su particularidad.

- El diseño físico debe estar orientado a generar un buen rendimiento en el procesamiento de consultas.

- Se deben definir también los procesos de administración futuros del *data warehouse*.

- El diseño físico inicial se deberá revisar de manera periódica para validar que continúa dando respuesta a las necesidades del cliente.

Una vez determinados qué hechos, dimensiones, métricas y atributos existen, podemos determinar también las claves foráneas que debe incluir el modelo físico. En este paso, es necesario tener en cuenta el tamaño adecuado de los atributos (por ejemplo, qué longitud tiene una cadena). También es relevante acordarse de crear las pertinentes claves primarias, claves foráneas y disparadores (por ejemplo, para actualizar de manera automática las claves primarias).

Para la tabla de hecho de hospitalización, tendremos:

	h_hosp	Tipo
(PK)	id_hosp	Clave primaria

	h_hosp	Tipo
(FK)	id_d_f_ingreso, id_d_f_alta, id_d_t_adm, id_d_t_adm_posterior, id_d_t_alta, id_d_f_urg_post, id_d_f_hosp_ant, id_d_f_hosp_post, id_d_GRD, id_d_servicio, id_d_diagnostico, id_d_sexo, id_d_tipo_ep	Claves foráneas
	d_hosp, est_ReaPQ, est_sep, est_UCI, est_UCI_CCA, est_uni_coro, num_unid, p_GRD, t_hosp	Métricas
	Episodio NHC	Atributos

Un detalle importante es que nuestra tabla de hechos tiene cinco referencias temporales. En lugar de crear cinco dimensiones temporales distintas, es mucho más lógico crear una sola dimensión temporal en un ámbito físico y que la tabla de hechos tenga cuatro claves foráneas que apuntan a esta dimensión, puesto que todas las dimensiones tienen el mismo nivel de profundidad. A esta única dimensión temporal la denominaremos **d_fecha**.

Dimensiones	Atributos
d_fecha	Año, mes, semana, día, hora, minuto, segundo

Otro detalle importante es que la dimensión **d_t_adm** y **d_t_adm_posterior** son la misma dimensión, por lo que podemos hacer el mismo juego anterior.

Esto simplifica el modelo final con el que trabajaremos, al tener menos dimensiones, y simplificará también los procesos de carga de dimensiones.

Tras esta reflexión, ha llegado el momento de pasar al diseño físico mediante Oracle XE. Se asume que el estudiante tiene a su disposición el usuario *system* (y su correspondiente clave). Como punto de partida, crearemos un usuario en Oracle para hacer esta parte y las siguientes. Por ejemplo, creamos el usuario DW_HOS con contraseña DW_HOS:

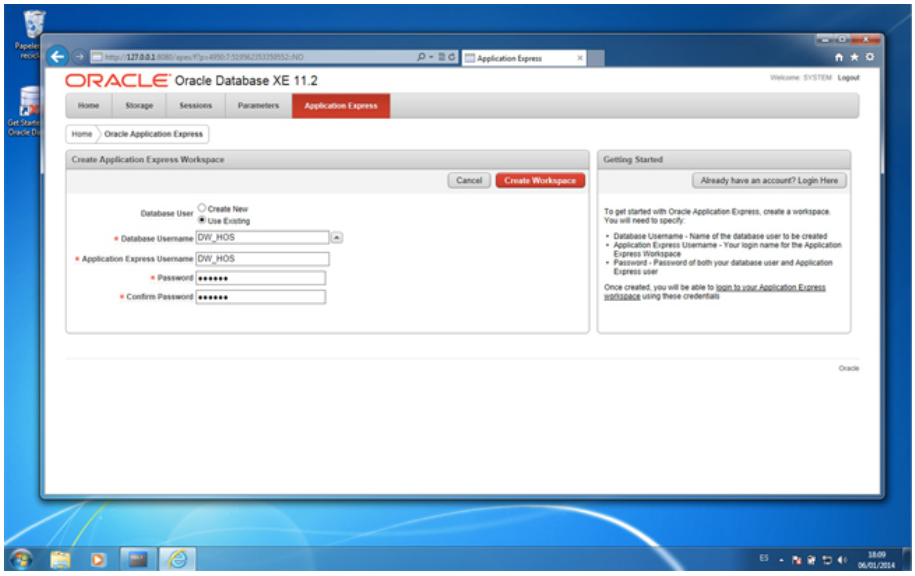
```
CREATE USER DW_HOS IDENTIFIED BY DW_HOS;
```

Y también proporcionamos permisos adecuados a este usuario:

```
GRANT ALL PRIVILEGES TO DW_HOS;
```

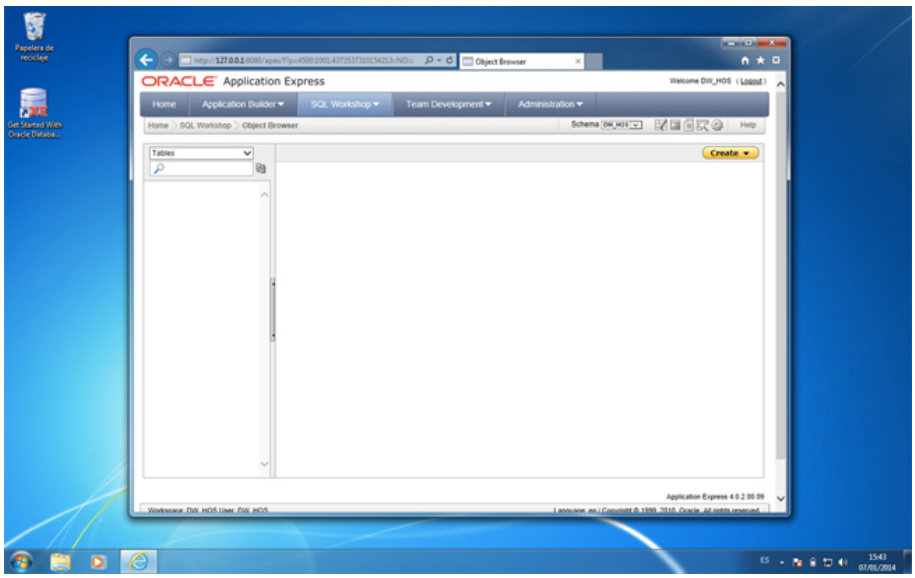
```
GRANT DBA TO DW_HOS;
```

Para crear las tablas mediante Oracle Application Express (Oracle APEX), es necesario crear un espacio de trabajo (*workspace*). Por lo tanto, como siguiente paso, creamos un *workspace* mediante Oracle APEX. Nos conectamos a APEX (de nuevo con el usuario *system* o con el usuario DW_HOS) y creamos un nuevo *workspace* para el usuario DW_HOS llamado de la misma manera.



Una vez conectados como el usuario DW_HOS en su *workspace* correspondiente, es posible proceder al diseño físico, es decir, a la creación de las tablas necesarias mediante SQL Workshop > Object Browser.

Para mejorar la visibilidad de las tablas con las que trabajamos, se han borrado las tablas por defecto que crea el sistema.



El proceso de creación de tablas es el mismo para todas las que vamos a crear.

1) Se pulsa el botón *Create*.

2) Se completa el nombre de la tabla.

3) Se añaden la clave primaria y los diferentes atributos de la dimensión y sus características.

4) Se añade la secuencia asociada a la clave primaria.

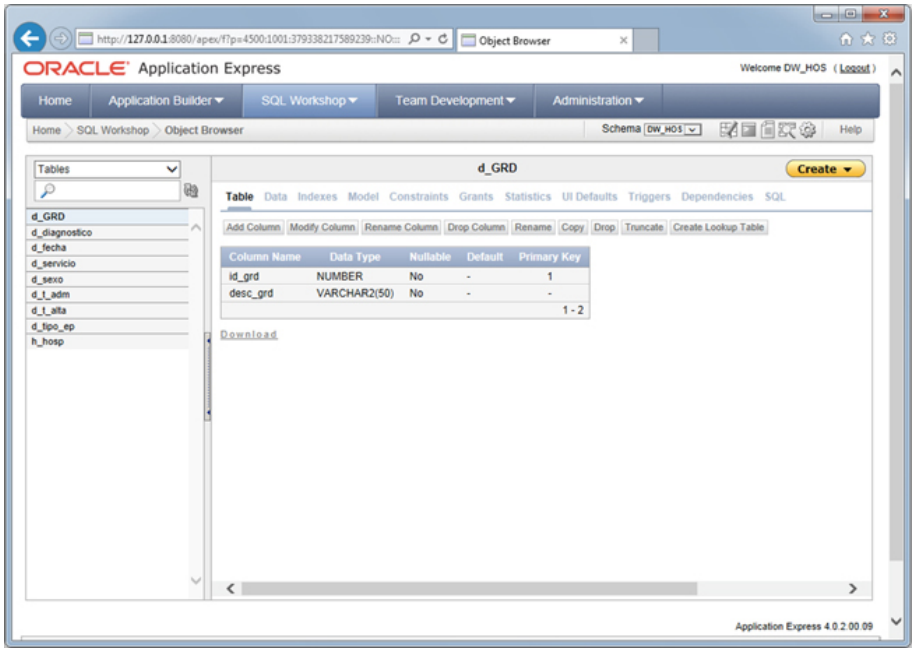
5) Se añaden las claves foráneas.

Como es lógico, primero se crean las tablas de dimensiones y por último, la tabla de hechos. De este modo, creamos cada una de las tablas de nuestro almacén de datos.

d_grd

Esta dimensión tiene tres campos.

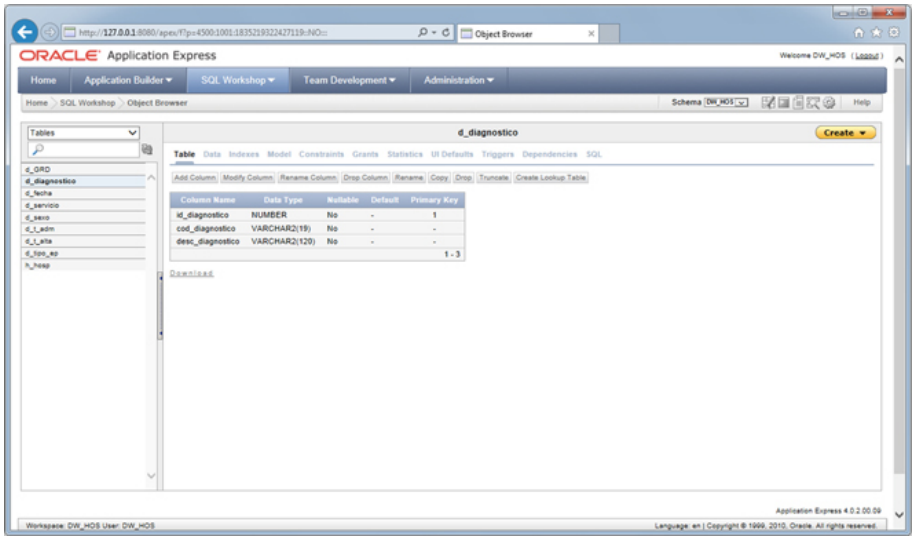
- id_grd: clave primaria, number
- cod_grd: código interno grd, varchar2(10)
- desc_grd: descripción, varchar2(50)



d_diagnostico

Esta dimensión tiene tres campos.

- id_diagnostico: clave primaria, number
- cod_diagnostico: código alfanúmero del diagnóstico, varchar2(19)
- desc_diagnostico: descripción del diagnóstico, varchar2(120)

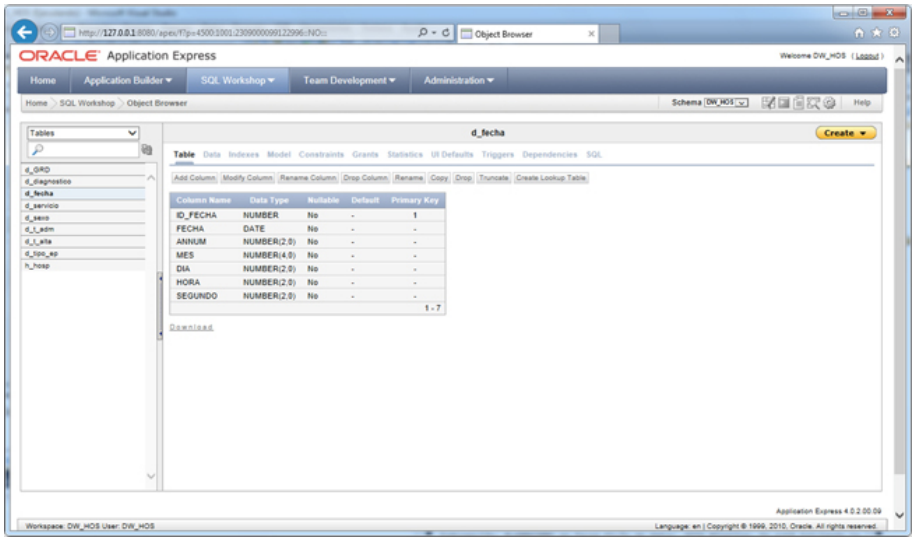


d_fecha

La dimensión temporal es la más compleja que existe. Se puede diseñar de muchas maneras distintas. Esta es simplemente una de las opciones, y podría ser mucho más completa. Esta dimensión tiene ocho campos.

- id_fecha: clave primaria, number
- fecha: fecha completa, date
- annum: año, number
- mes: mes, number
- día: día, number
- hora: hora, number
- segundo: segundo, number

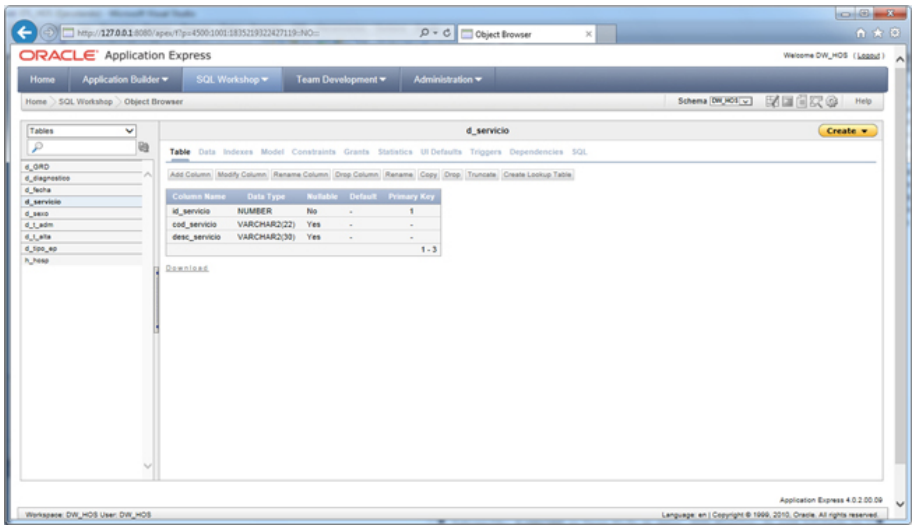
Un detalle respecto a esta dimensión es que, a partir de la fecha completa que nos proporcionan los datos, derivaremos –es decir, calcularemos automáticamente– el resto de los campos de la dimensión. Este proceso se llevará a cabo en un ámbito de ETL.



d_servicio

Esta dimensión tiene tres campos.

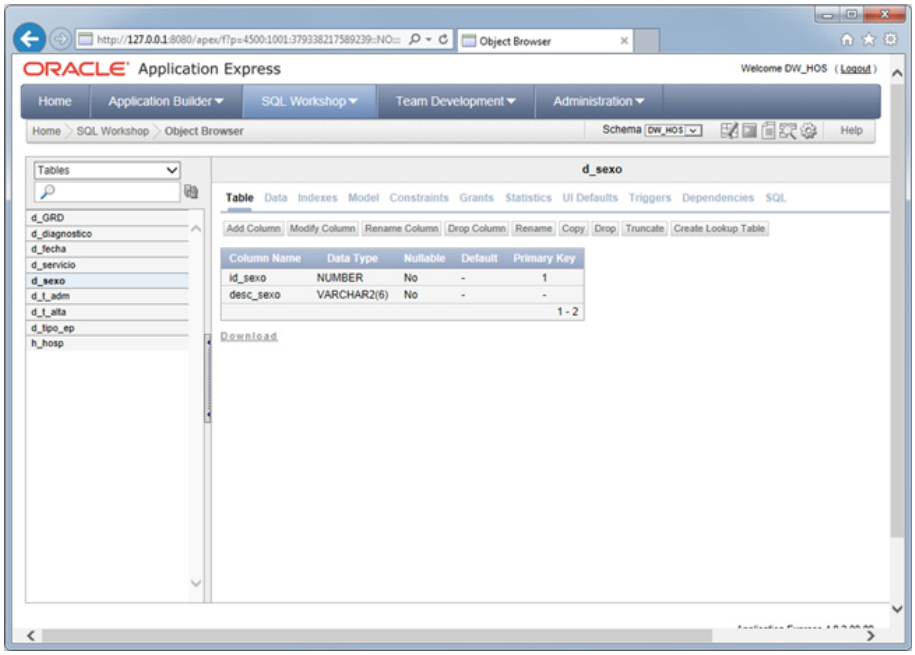
- id_servicio: clave primaria, number
- cod_servicio: código del servicio, varchar2(22)
- desc_servicio: descripción del servicio, varchar2(30)



d_sexo

Esta dimensión tiene dos campos.

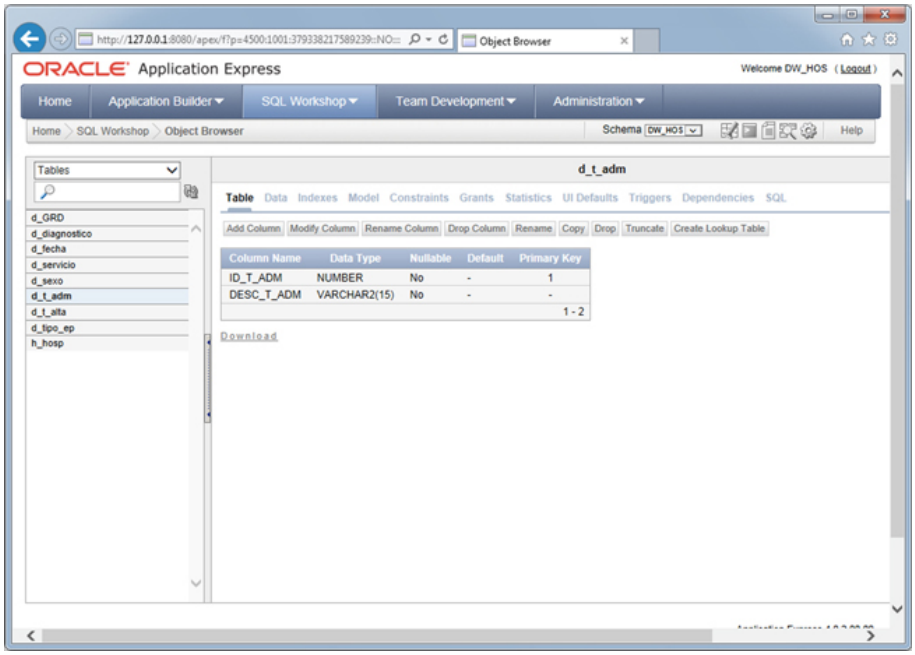
- id_sexo: clave primaria, number
- desc_sexo: descripción, varchar2(6)



d_t_adm

Esta dimensión tiene dos campos.

- id_t_adm: clave primaria, number
- desc_t_adm: descripción del tipo de admisión, varchar2(15)

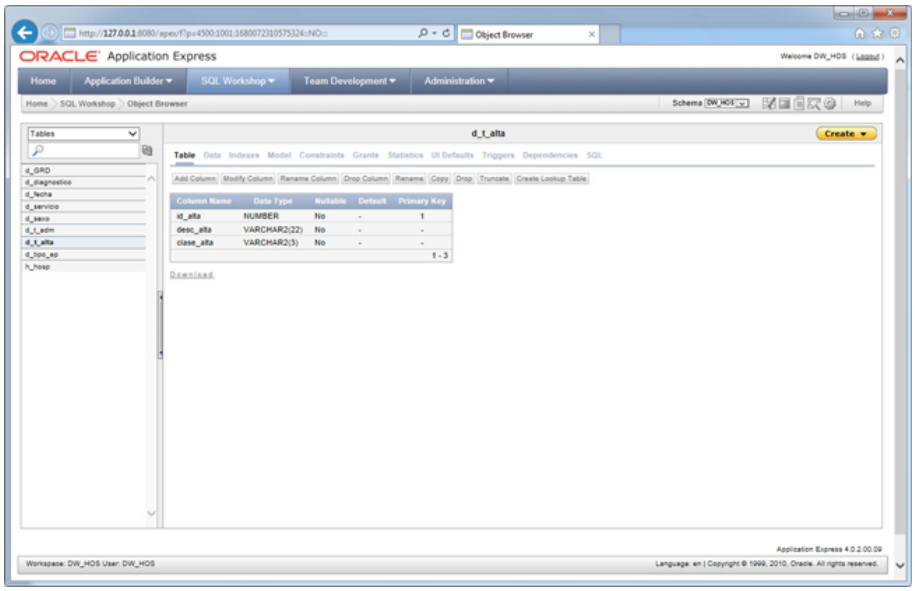


d_t_alta

Esta dimensión tiene tres campos.

- id_alta: clave primaria, number
- desc_alta: descripción del tipo de alta, varchar2(22).

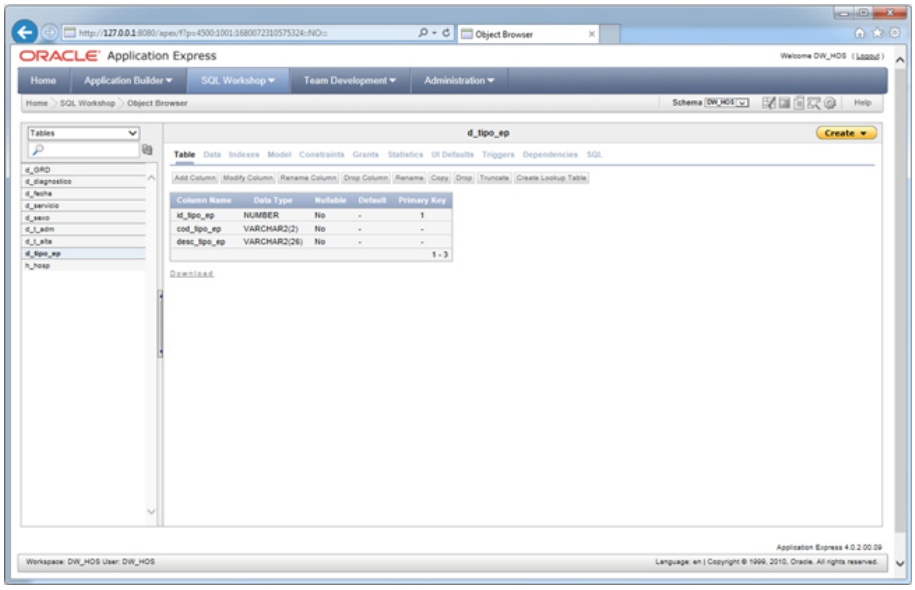
- clase_alta: código de clasificación del alta, varchar2(3)



d_tipo_ep

Esta dimensión tiene tres campos.

- id_tipo_ep: clave primaria, number
- cod_tipo_ep: código, varchar2(2)
- desc_ tipo_ep: descripción del tipo de episodio, varchar2(26)

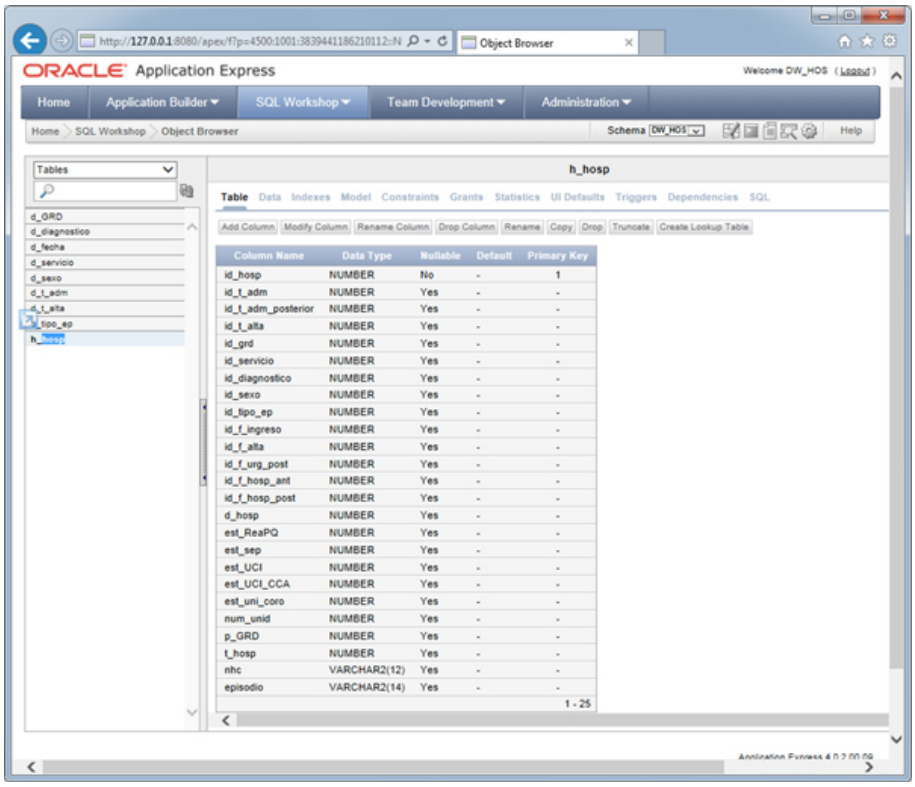


h_hosp

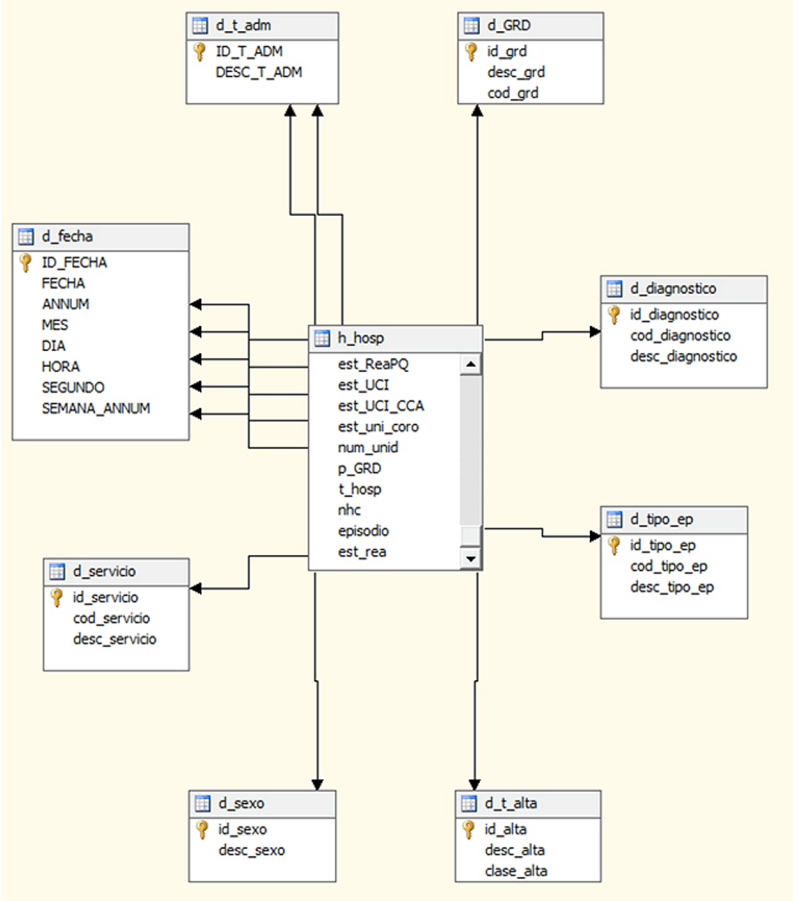
La tabla de hecho está compuesta por lo siguiente.

- id_hosp: clave primaria
- Las 13 claves foráneas que se comentan anteriormente: number

- 9 métricas: number
- 2 atributos:
  - nhc: varchar2(12)
  - Episodio: varchar2 (14)



El esquema resultante, teniendo en cuenta el comentario de la dimensión temporal, es el siguiente:



Por lo que tenemos una tabla de hecho y ocho tablas de dimensiones. Como ya hemos comentado, las ocho dimensiones en realidad representan trece dimensiones (como podemos validar con las relaciones entre la tabla de hecho y las dimensiones).

A continuación, se encuentra el *script* de creación de las tablas de la base de datos.

```
CREATE TABLE "d_GRD"
(
"id_grd" NUMBER NOT NULL ENABLE,
"cod_grd" VARCHAR2(10) NOT NULL ENABLE,
"desc_grd" VARCHAR2(60) NOT NULL ENABLE,
CONSTRAINT "d_GRD_PK" PRIMARY KEY ("id_grd") ENABLE
) ;
```

```
CREATE TABLE "d_diagnostico"
(
"id_diagnostico" NUMBER NOT NULL ENABLE,
"cod_diagnostico" VARCHAR2(19) NOT NULL ENABLE,
"desc_diagnostico" VARCHAR2(120) NOT NULL ENABLE,
CONSTRAINT "d_diagnostico_PK" PRIMARY KEY ("id_diagnostico") ENABLE
```

```
) ;
```

```
CREATE TABLE "d_fecha"
(
"FECHA" DATE NOT NULL ENABLE,
"ANNUM" NUMBER(2,0) NOT NULL ENABLE,
"MES" NUMBER(4,0) NOT NULL ENABLE,
"DIA" NUMBER(2,0) NOT NULL ENABLE,
"HORA" NUMBER(2,0) NOT NULL ENABLE,
"SEGUNDO" NUMBER(2,0) NOT NULL ENABLE,
"ID_FECHA" NUMBER NOT NULL ENABLE,
CONSTRAINT "D_FECHA_PK" PRIMARY KEY ("ID_FECHA") ENABLE
) ;
```

```
CREATE TABLE "d_servicio"
(
"id_servicio" NUMBER NOT NULL ENABLE,
"cod_servicio" VARCHAR2(22) NOT NULL ENABLE,
"desc_servicio" VARCHAR2(30) NOT NULL ENABLE,
CONSTRAINT "d_servicio_PK" PRIMARY KEY ("id_servicio") ENABLE
) ;
```

```
CREATE TABLE "d_sexo"
(
"id_sexo" NUMBER NOT NULL ENABLE,
"desc_sexo" VARCHAR2(6) NOT NULL ENABLE,
CONSTRAINT "d_sexo_PK" PRIMARY KEY ("id_sexo") ENABLE
) ;
```

```
CREATE TABLE "d_t_adm"
(
"ID_T_ADM" NUMBER NOT NULL ENABLE,
"DESC_T_ADM" VARCHAR2(15) NOT NULL ENABLE,
CONSTRAINT "D_T_ADM_PK" PRIMARY KEY ("ID_T_ADM") ENABLE
) ;
```

```
CREATE TABLE "d_t_alta"
(
"id_alta" NUMBER NOT NULL ENABLE,
"desc_alta" VARCHAR2(22) NOT NULL ENABLE,
"clase_alta" VARCHAR2(3) NOT NULL ENABLE,
CONSTRAINT "d_alta_PK" PRIMARY KEY ("id_alta") ENABLE
) ;
```

```
CREATE TABLE "d_tipo_ep"
( "id_tipo_ep" NUMBER NOT NULL ENABLE,
"cod_tipo_ep" VARCHAR2(2) NOT NULL ENABLE,
```

```
"desc_tipo_ep" VARCHAR2(26) NOT NULL ENABLE,
CONSTRAINT "d_tipo_ep_PK" PRIMARY KEY ("id_tipo_ep") ENABLE
) ;
```

3.4.1. Metadatos

En cada fase del diseño de una factoría de información corporativa, debemos tener en cuenta los metadatos necesarios y asociados. Aunque trabajamos con un ejemplo limitado, podemos reflexionar sobre los metadatos que hay que considerar. En un ámbito de diseño del almacén de datos, es necesario tener en cuenta los metadatos siguientes.

- Sobre la factoría de información:
  - Arquitectura de la factoría de información
  - Tipo de tablas de hecho en el almacén de datos
  - Tipos de dimensiones en el almacén de datos
  - Información sobre la base de datos

- Sobre las tablas:
  - Nombre, atributos y rol

- Sobre las columnas:
  - Nombre, tipo de dato, longitud, reglas de edición, definición

Parte de esta información está presente en este documento, y también es accesible por medio del sistema gestor del almacén de datos. Frecuentemente, se incluye un gestor de metadatos para gestionar de manera correcta y en único punto este tipo de información y favorecer posteriores desarrollos.

Un tipo de metadato más vinculado con el ciclo de vida del dato describe cuándo es creado y cuándo, de ser necesario, es modificado. Aunque no lo hemos contemplado en nuestra solución, por su limitación de alcance, cada tabla puede incluir un par de campos para gestionar este tipo de información.

3.4.2. Optimización de la factoría de información

Este ejemplo es pequeño, por lo que parte de lo que vamos a comentar en esta sección es opcional. La optimización de la factoría de información es un proceso continuo. Hay que tener en cuenta:

- La previsión de crecimiento de datos.
- La previsión del tipo de consultas que se harán.

Con este tipo de información, se implementarían:

- Propuesta de *tablespace* para Oracle.
- Índices extras para mejorar las consultas.

- Vistas materializadas

Después de la carga de datos, discutiremos sobre el tamaño actual del almacén y sus necesidades futuras.

Una vez que el *data warehouse* está diseñado e implementado en la base de datos, el siguiente paso es la creación de los procesos de carga de datos en el mismo, tal y como se ven en el apartado siguiente.

4. Carga de datos

Una vez tenemos una propuesta de diseño del almacén de datos, y ha sido implementada en la base de datos con la que trabajamos, es el momento de centrarnos en los procesos ETL. Como ya sabemos, estos procesos consisten en la extracción, transformación y carga de los datos. En definitiva, lo que se persigue es estructurar y acomodar los datos de las fuentes de origen en el almacén de datos.

En nuestro caso particular, tenemos solo una fuente de origen, un fichero Excel, y una fuente de destino, Oracle, la base de datos donde se aloja nuestro almacén de datos.

4.1. Identificación de los procesos ETL necesarios

Los procesos ETL deben conceptualizarse como manipulaciones de flujos de datos. Estos procesos deben diseñarse teniendo en cuenta distintos factores como los siguientes.

- Cómo debe cargarse de manera lógica la información, es decir, qué debe cargarse primero y qué después.

- La ventana de tiempo disponible, hecho que puede condicionar lo que debemos cargar.

- Tipo de carga: inicial o incremental.

En nuestro caso, está claro que esta es una carga inicial, por lo que nuestros procesos no se diseñarán con el objetivo de repetirse de manera periódica. También es cierto que desconocemos la ventana de tiempo disponible (y no es una condición para esta actividad), pero en el contexto de producción, este es un factor muy relevante. Un diseño incorrecto de los procesos ETL puede significar que la información no está disponible en el almacén de datos y, por lo tanto, que no está disponible para usuarios, procesos y aplicaciones que se apalancan sobre esta información de calidad.

El primer paso en la creación de los procesos ETL es comprender qué procesos son necesarios y en qué orden deben llevarse a cabo. En general, el criterio de los procesos ETL sigue estas pautas:

- Se identifica si los datos se deben cargar en un área intermedia. En nuestro caso, no es necesario.

- Podemos diferenciar dos tipos de situaciones en nuestro caso particular.

  - Dimensiones con valores fijos ya conocidos, presentes en Excel en la hoja de descripción de tabla y que no van a cambiar en el tiempo. Estas dimensiones son: sexo, tipo de episodio, tipo de admisión y tipo de alta. Estos valores son reducidos y se insertarán directamente en la base de datos.

  - Dimensiones con valores no fijos que se extraerán, transformarán y cargarán mediante procesos ETL. Estas dimensiones son: fecha, diagnóstico y servicio.

- Por último, la tabla de hecho, hospitalización, que también se cargará mediante un proceso ETL.

- Las dimensiones se cargan antes que las tablas de hecho.

Por lo tanto, y teniendo en cuenta nuestro caso particular, identificamos dos conjuntos de tareas:

- Proceso de creación de las dimensiones estáticas que llevaremos a cabo directamente desde Oracle APEX.

- Procesos ETL que primero cargan las dimensiones dinámicas fecha, diagnóstico, GRD y servicio y, después, la tabla de hecho de hospitalización. Este será un único proceso ETL.

Por lo tanto, hemos identificado cinco procesos ETL para crear.

4.2. Descripción de las acciones en cada proceso ETL

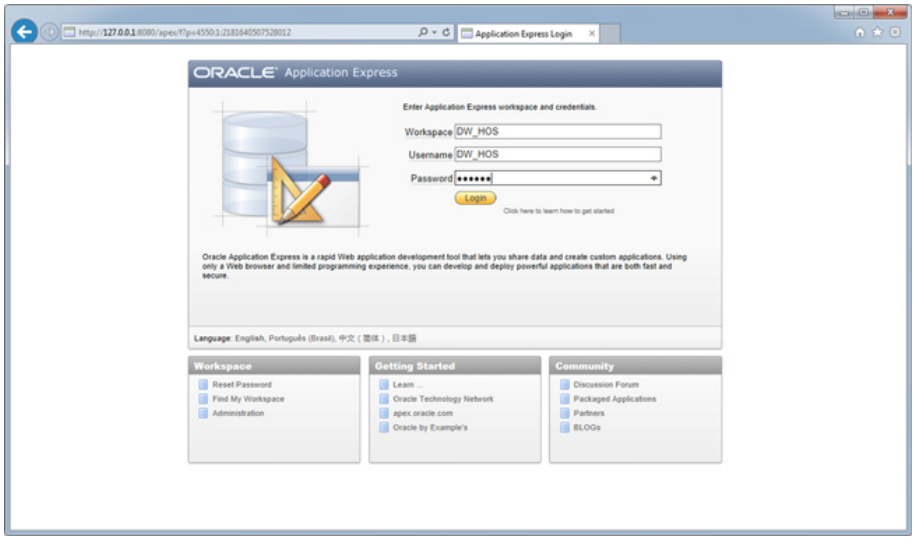
Vamos a describir en este apartado la funcionalidad a alto nivel de los procesos ETL que hemos identificado. Tener claro a alto nivel qué debe hacer el ETL ayuda *a posteriori* en el proceso de diseño, usando una herramienta específica.

Proceso ETL	Descripción	Subtareas
Carga dimensiones estáticas	Carga de las dimensiones sexo, tipo de episodio, tipo de admisión y tipo de alta. Este proceso no se encapsulará mediante un proceso ETL, sino directamente en la base de datos.	Insertar valores dimensión sexo. Insertar valores dimensión tipo de episodio. Insertar valores dimensión tipo de admisión. Insertar valores tipo de alta.
Carga dimensión diagnostico	Carga de la dimensión diagnóstico por medio de un proceso ETL.	Extraer valores Excel. Omitir campos innecesarios. Renombrar nombres valores. Ordenar y quedarse con únicos. Cargar los datos en la tabla correspondiente.

Proceso ETL	Descripción	Subtareas
Carga dimensión servicio	Carga de la dimensión servicio por medio de un proceso ETL.	Extraer valores Excel. Omitir campos innecesarios. Renombrar nombres valores. Ordenar y quedarse con únicos. Cargar los datos en la tabla correspondiente.
Carga dimensión GRD	Carga de la dimensión GRD por medio de un proceso ETL.	Extraer valores Excel. Omitir campos innecesarios. Renombrar nombres valores. Ordenar y quedarse con únicos. Cargar los datos en la tabla correspondiente.
Carga dimensión fecha	Carga de la dimensión fecha por medio de un proceso ETL.	Extraer valores Excel. Omitir campos innecesarios. Renombrar nombres valores. Ordenar y quedarse con únicos. Cargar los datos en la tabla correspondiente.
Carga tabla de hechos	Carga de la tabla de hechos hospitalización por medio de un proceso ETL.	Extraer valores Excel. Omitir campos innecesarios. Renombrar nombres valores. Recuperar id para cada una de las dimensiones. Cargar tabla de hecho.

4.2.1. Diseño procesos ETL

Vamos a empezar con la carga de las dimensiones estáticas. Entremos en Oracle Apex, con el usuario DW_HOS.



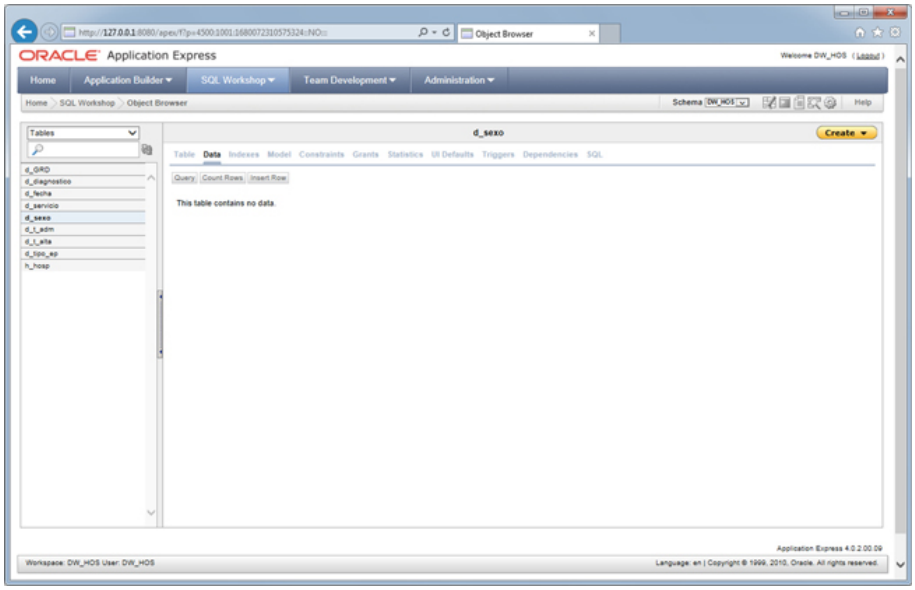
Una vez dentro, volvemos a la sección SQL Workshop > Object Browser.

Empezamos por la dimensión sexo. En el fichero Excel, tenemos los dos valores que hay que insertar:

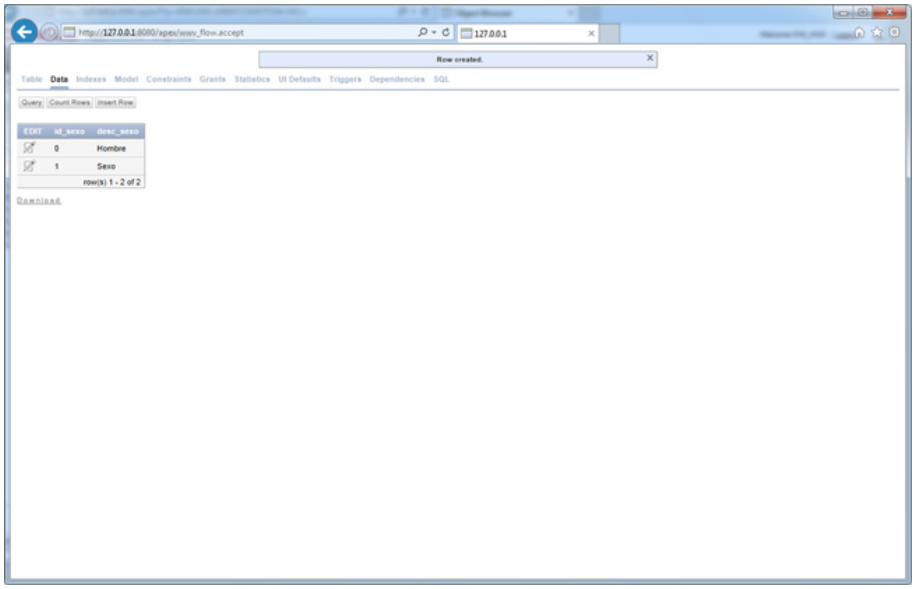
- 0 - Hombre

- 1 - Mujer

Seleccionamos la dimensión d_sexo, sección *data*. Tenemos la opción *Insert row* para insertar manualmente estos valores.



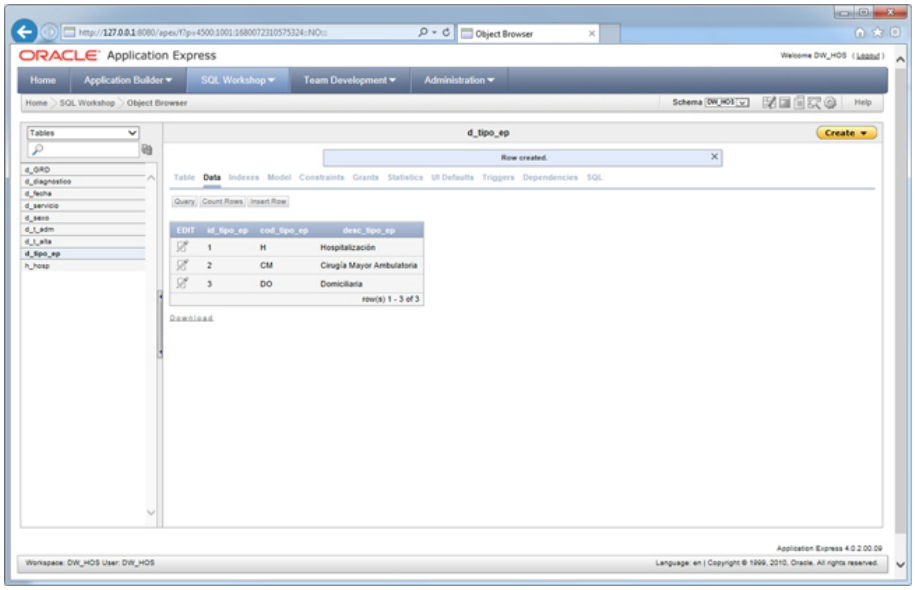
Tras la inserción, considerando los valores anteriores, tenemos:



Continuamos con la dimensión tipo de episodio. En el fichero Excel, tenemos los tres valores que hay que insertar:

- H - Hospitalización
- CM - Cirugía mayor ambulatoria
- DO - Domiciliaria

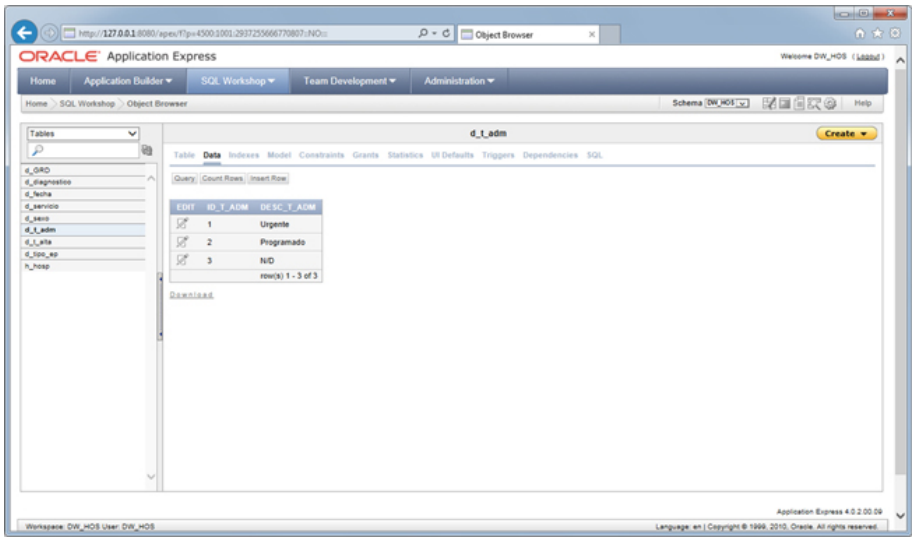
Procedemos como en el caso anterior, y el resultado es:



Continuamos con la dimensión tipo de admisión. En el fichero Excel, tenemos los tres valores que hay que insertar:

- 1 - Urgente
- 2 - Programado
- 9 - N/D

Procedemos como en el caso anterior, y el resultado es:

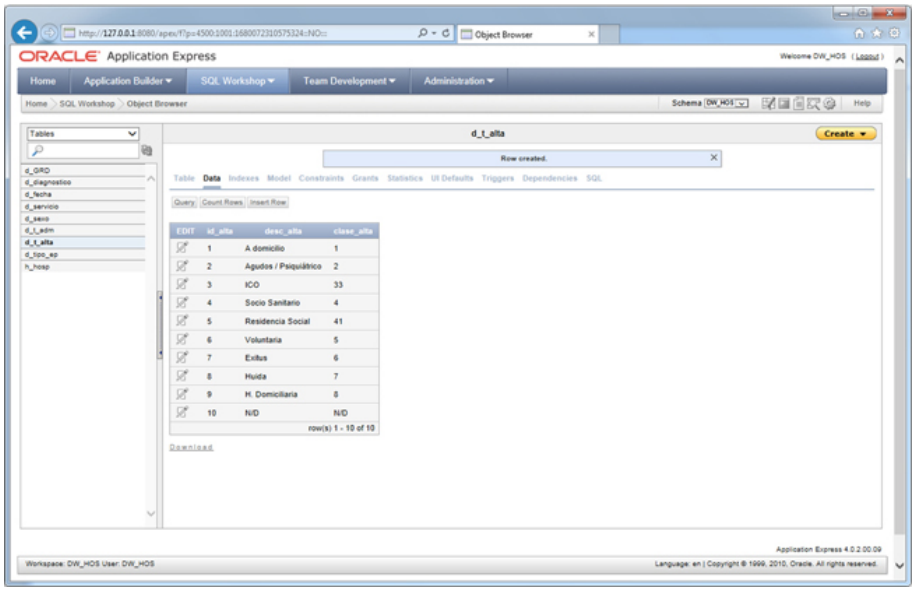


Finalmente, procedemos con la dimensión tipo de alta. En el fichero Excel, tenemos los diez valores que hay que insertar (clase alta y descripción alta):

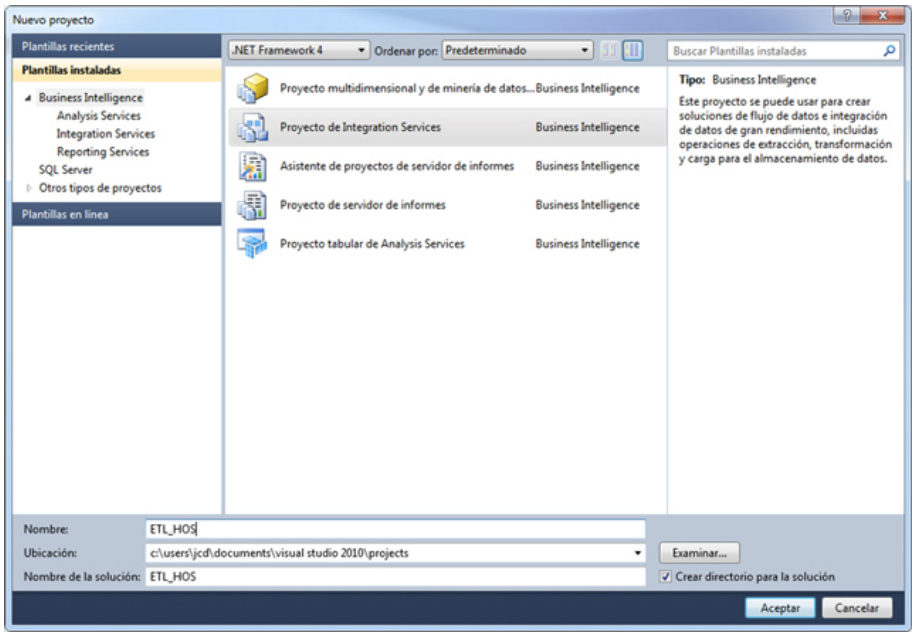
- 1 - A domicilio
- 2 - Agudos/psiquiátrico
- 33 - ICO
- 4 - Sociosanitario
- 41 - Residencia social

- 5 - Voluntaria
- 6 - *Exitus*
- 7 - Huida
- 8 - H. domiciliaria
- N/D - N/D

Procedemos como en el caso anterior, y el resultado es:



El diseño de los procesos ETL se hará con Microsoft Visual Studio e Integration Services. Iniciamos un nuevo proyecto y creamos un nuevo proyecto de Integration Services.



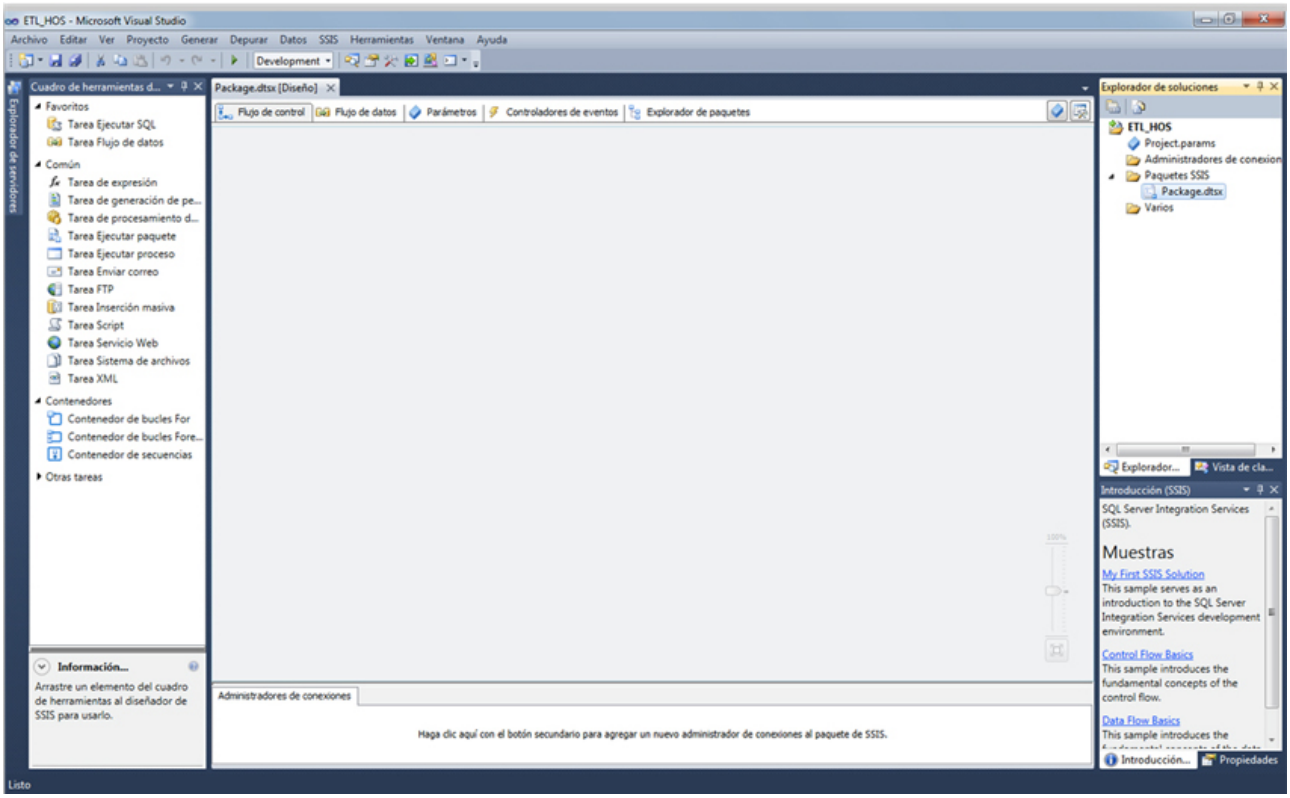
El entorno de trabajo de Integration Services tiene diferentes zonas de trabajo.

- Cuadro de herramientas: tareas ETL disponibles (en la parte izquierda de la pantalla).

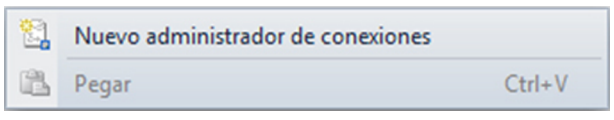
- Explorador de soluciones: estructura del proyecto (en la parte superior derecha de la pantalla).

- Explorador de elemento: que muestra las propiedades del mismo (en la parte inferior derecha de la pantalla).

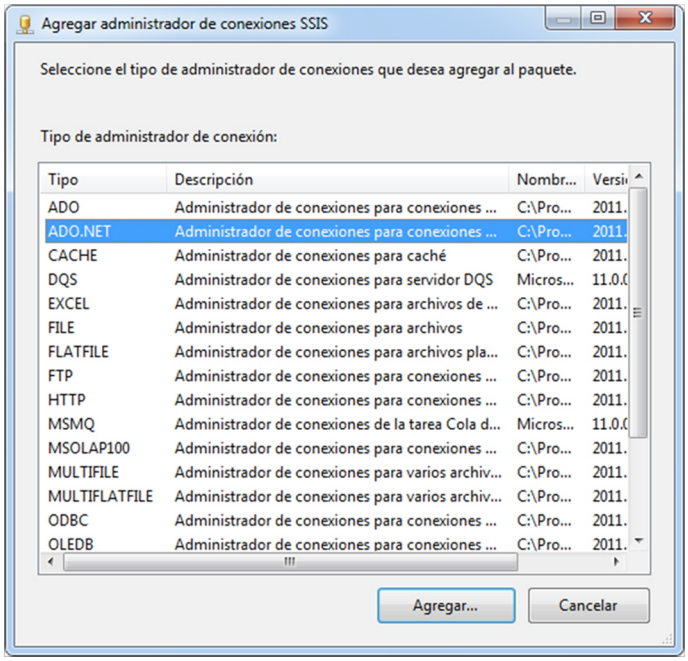
- Área de trabajo: en la que se definen los paquetes de ETL (en la parte central de la pantalla). Tenemos diferentes opciones: flujo de control, flujo de datos, parámetros, controladores de eventos y explorador de paquetes.



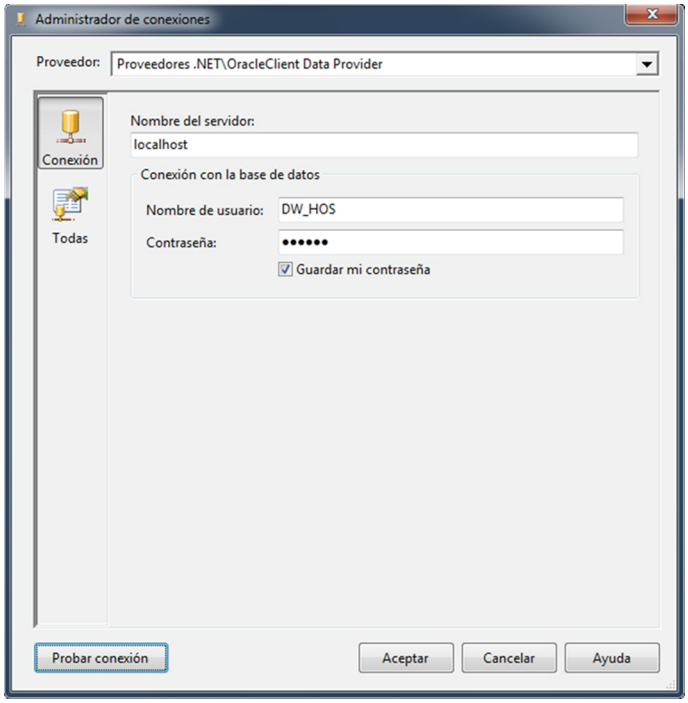
Nuestro proceso ETL tiene una fuente de origen (Excel) y una fuente de destino (Oracle). Para que estén disponibles para todo el proyecto, en el Explorador de soluciones > Administración de conexiones, añadimos las dos.



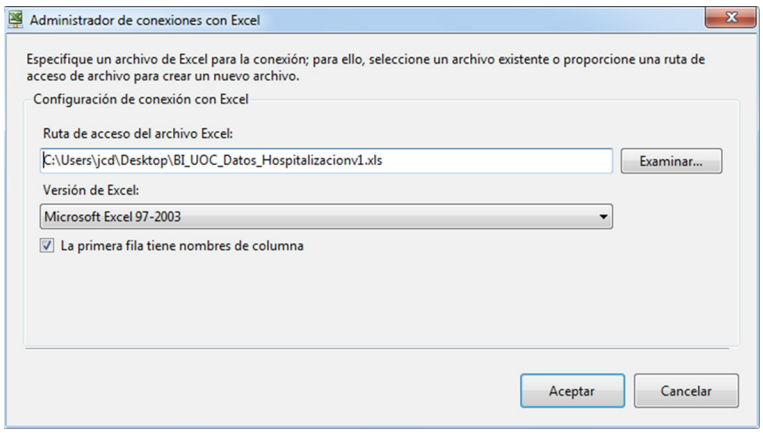
Para crear la conexión con Oracle, seleccionamos ADO.NET.



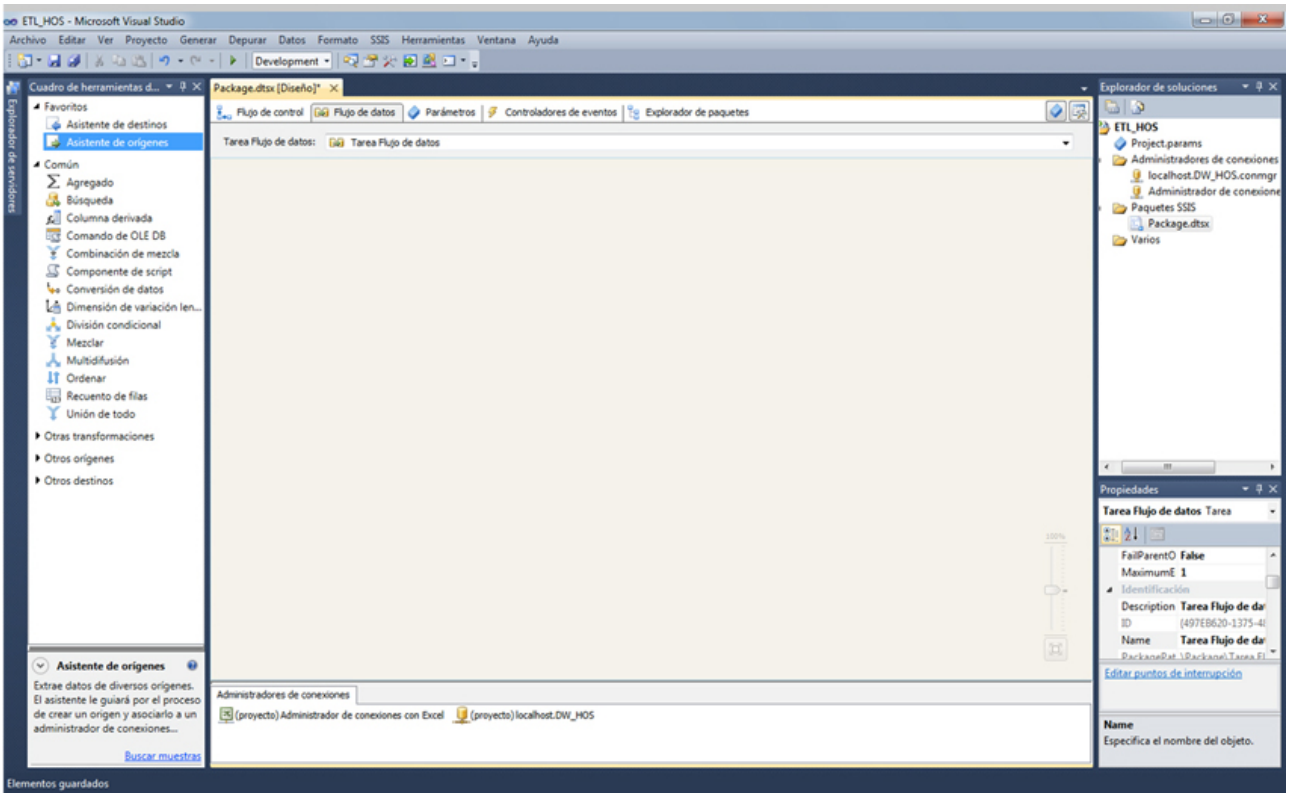
En este menú, seleccionamos Proveedor.NET\OracleClient Data Provider, completamos con el usuario y contraseña DW_HOS y guardamos la contraseña.



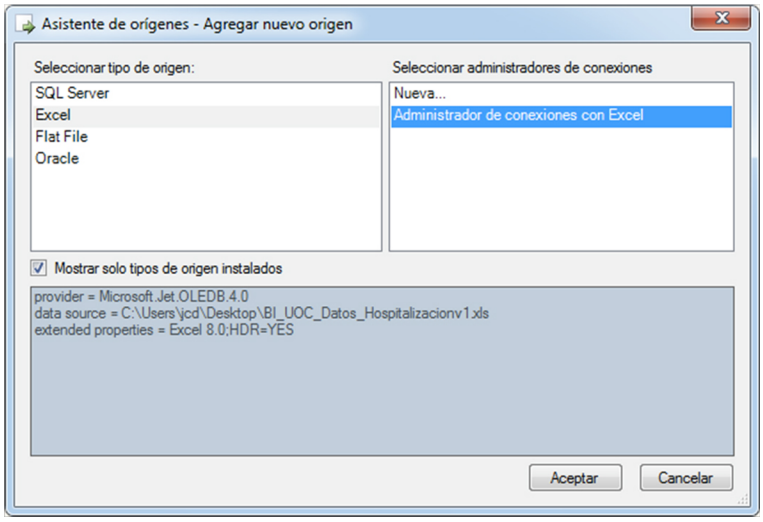
Para crear la conexión con el fichero Excel, añadimos una nueva conexión y seleccionamos tipo Excel. En esta solución, el fichero Excel está en el escritorio. Debemos indicar la ruta del fichero en el menú siguiente.



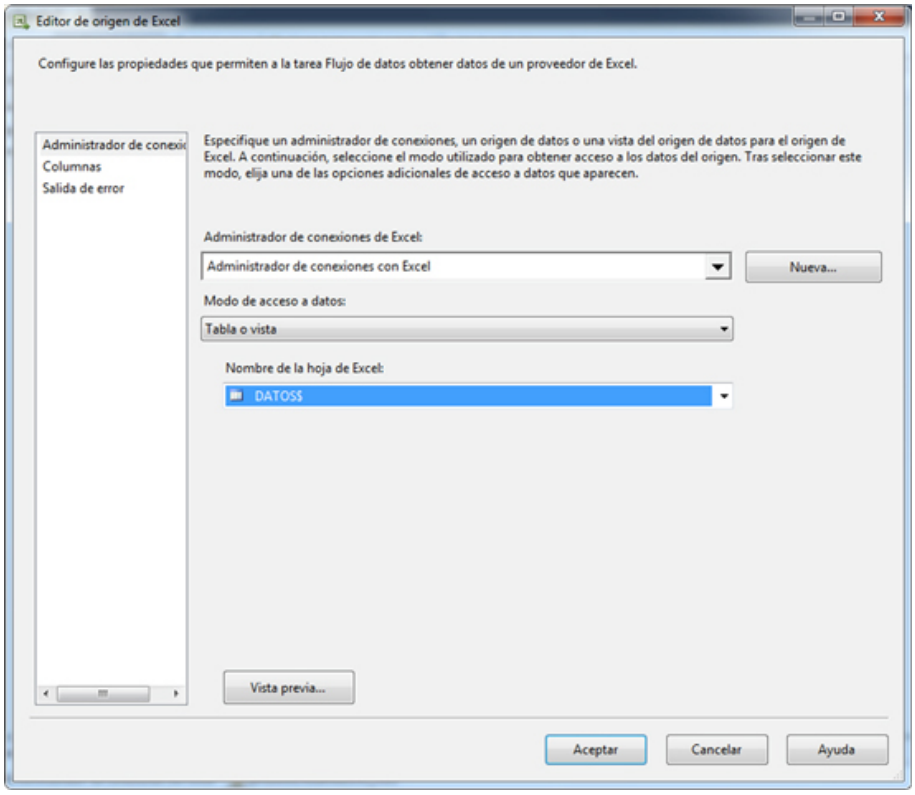
Ahora ya estamos preparados para crear nuestro proceso ETL. El primer paso es extraer los registros del fichero Excel. Creamos una tarea de flujos de datos.



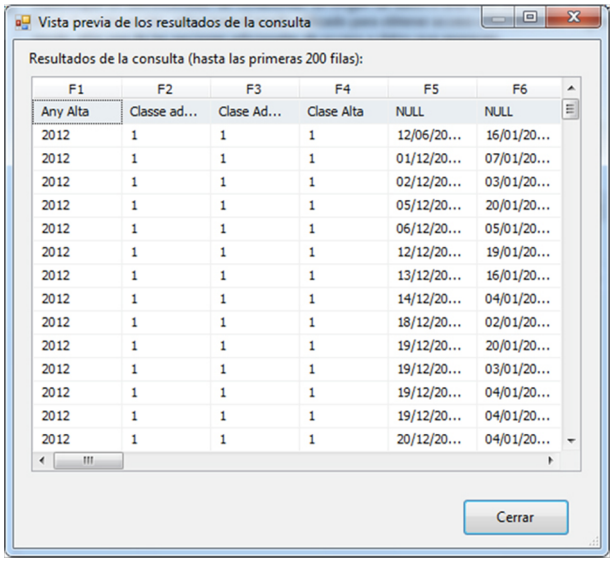
Agregamos el origen del flujo (en nuestro caso, Excel).



Pulsamos sobre el objeto creado en el área de trabajo y configuramos los parámetros de acceso a la hoja de Excel. Primero, a la hoja a la que accede el paso.



Validamos la conectividad.



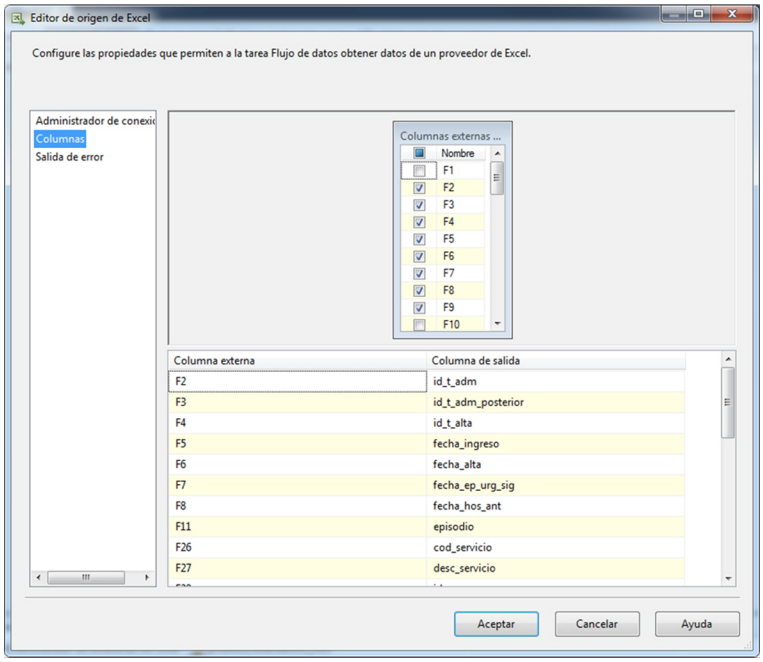
F1	F2	F3	F4	F5	F6
Any Alta	Classe ad...	Clase Ad...	Clase Alta	NULL	NULL
2012	1	1	1	12/06/20...	16/01/20...
2012	1	1	1	01/12/20...	07/01/20...
2012	1	1	1	02/12/20...	03/01/20...
2012	1	1	1	05/12/20...	20/01/20...
2012	1	1	1	06/12/20...	05/01/20...
2012	1	1	1	12/12/20...	19/01/20...
2012	1	1	1	13/12/20...	16/01/20...
2012	1	1	1	14/12/20...	04/01/20...
2012	1	1	1	18/12/20...	02/01/20...
2012	1	1	1	19/12/20...	20/01/20...
2012	1	1	1	19/12/20...	03/01/20...
2012	1	1	1	19/12/20...	04/01/20...
2012	1	1	1	19/12/20...	04/01/20...
2012	1	1	1	20/12/20...	04/01/20...

Ahora, debemos acabar de configurar el paso para hacer dos tareas al mismo tiempo.

- Primero, omitir aquellos campos que hemos identificado que son innecesarios. Para esto, deseleccionamos las columnas F1, F10, F12-F21 y F25.

- Renombramos el resto de las columnas o bien con el nombre del campo en la base de datos, o bien con un nombre que *a posteriori* nos permita identificar su contenido. De esta manera:
  - F2 será id_t_adm
  - F3 será id_t_adm_posterior
  - F4 será clase_alta
  - F5 será fecha_ingreso
  - F6 será fecha_alta
  - F7 será fecha_ep_urg_sig
  - F8 será fecha_hos_ant
  - F9 será fecha_hos_pos
  - F11 será episodio
  - F22 será cod_grd
  - F23 será desc_grd
  - F24 será nhc
  - F26 será cod_servicio
  - F27 será desc_servicio
  - F28 será id_sexo
  - F29 será cod_tipo_ep
  - F30 será d_hosp
  - F31 será est_ReaPQ
  - F32 será est_rea
  - F33 será est_UCI
  - F34 será est_UCI_CCA
  - F35 será est_uni_coro
  - F36 será num_unid

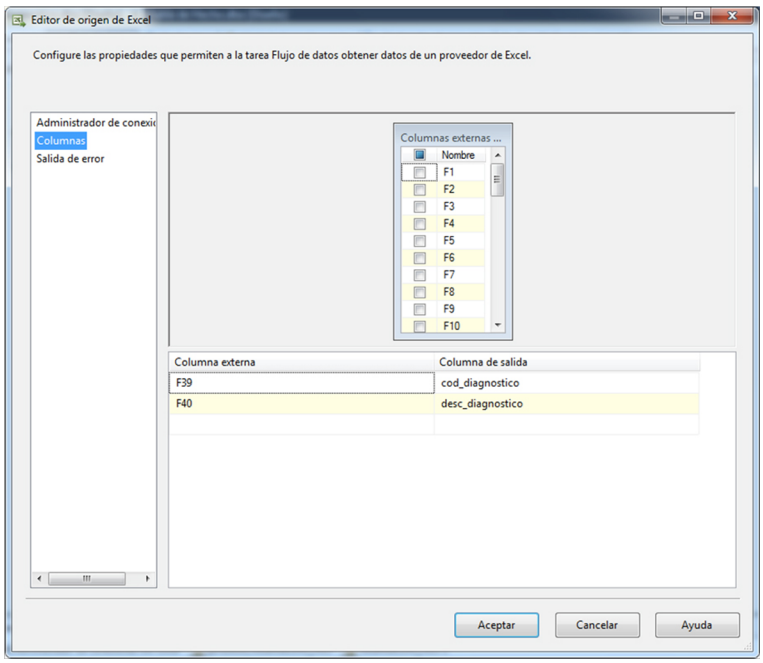
- F37 será p_GRD
- F38 será t_hosp
- F39 será cod_diagnostico
- F40 será desc_diagnostico



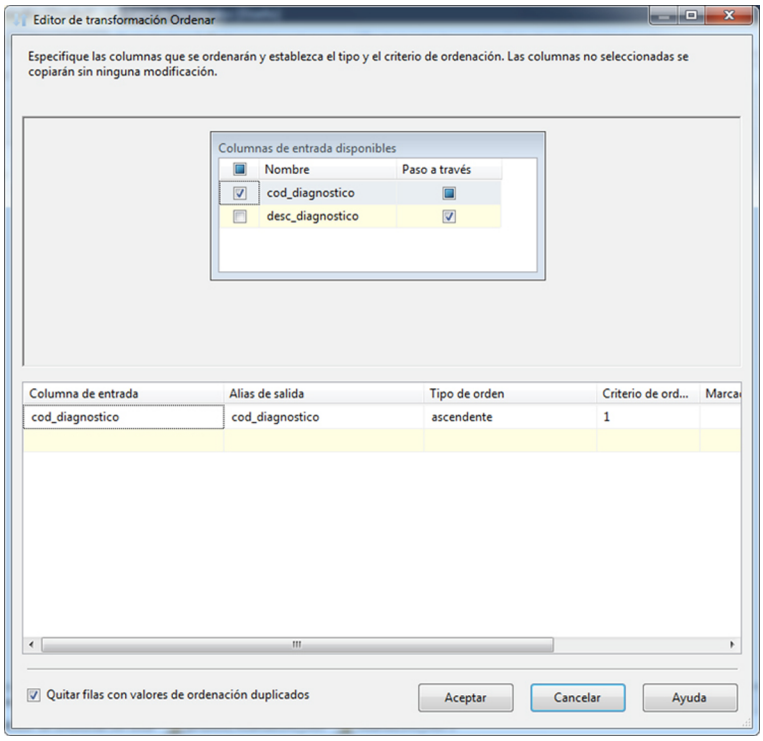
Vamos a usar paso configurado como base para todos los ETL, para cargar las dimensiones.

4.2.2. Dimensión diagnóstico

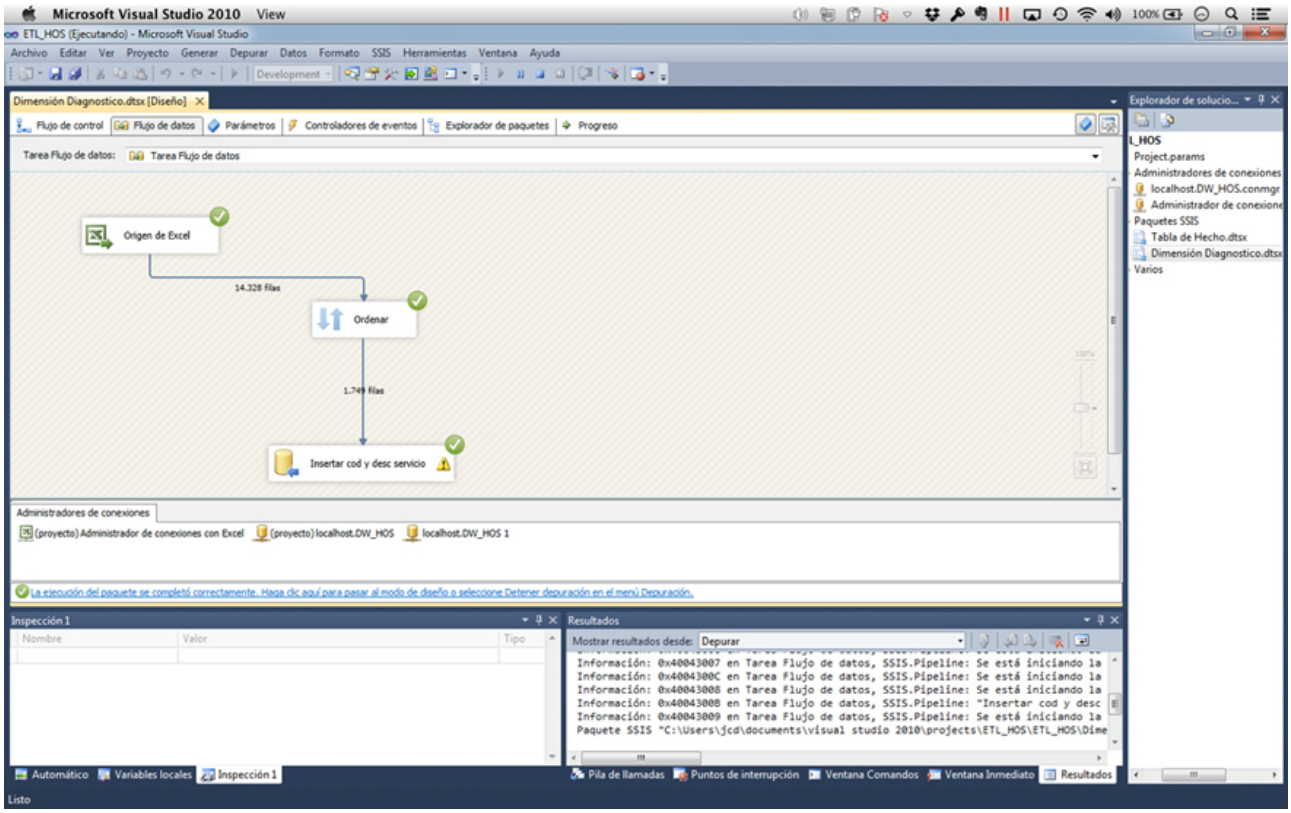
Creamos un nuevo *package* dentro del proyecto llamado dimensión diagnóstico. Copiamos el paso de Origen Excel que hemos creado. Borramos todas las columnas menos la 39 y 40, que hacen referencia a los datos de diagnóstico.



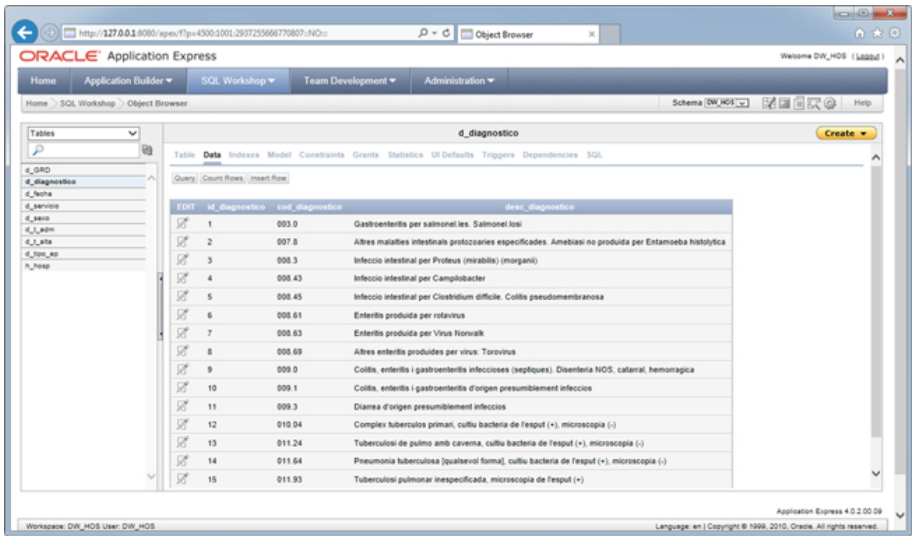
Añadimos un paso de ordenar y quitamos los duplicados.



Añadimos un paso de inserción, lo configuramos correctamente para que apunte a la dimensión correspondiente y ejecutamos este proceso ETL.



Como resultado, habremos insertado 1.747 registros.

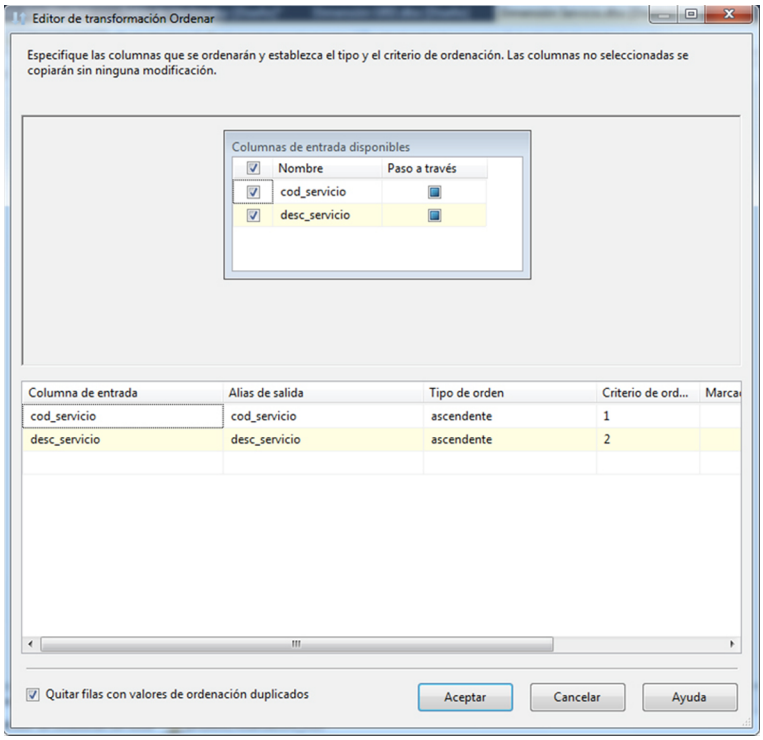


Durante el proceso de desarrollo, es posible hacer alguna carga de prueba. Después de la misma, será necesario truncar la tabla y resetear la secuencia de la tabla a cero antes de la inserción final.

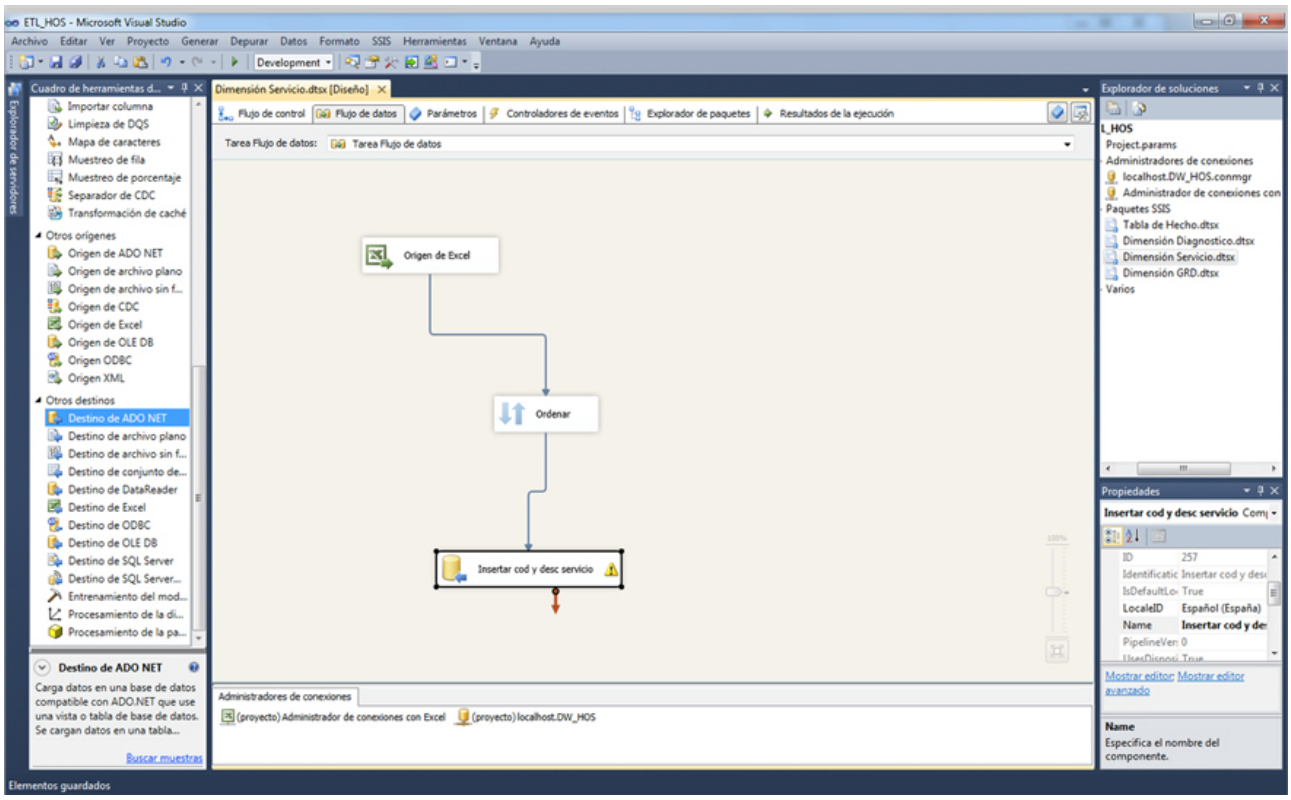
Este mismo proceso se hace para servicio y GRD.

4.2.3. Dimensión servicio

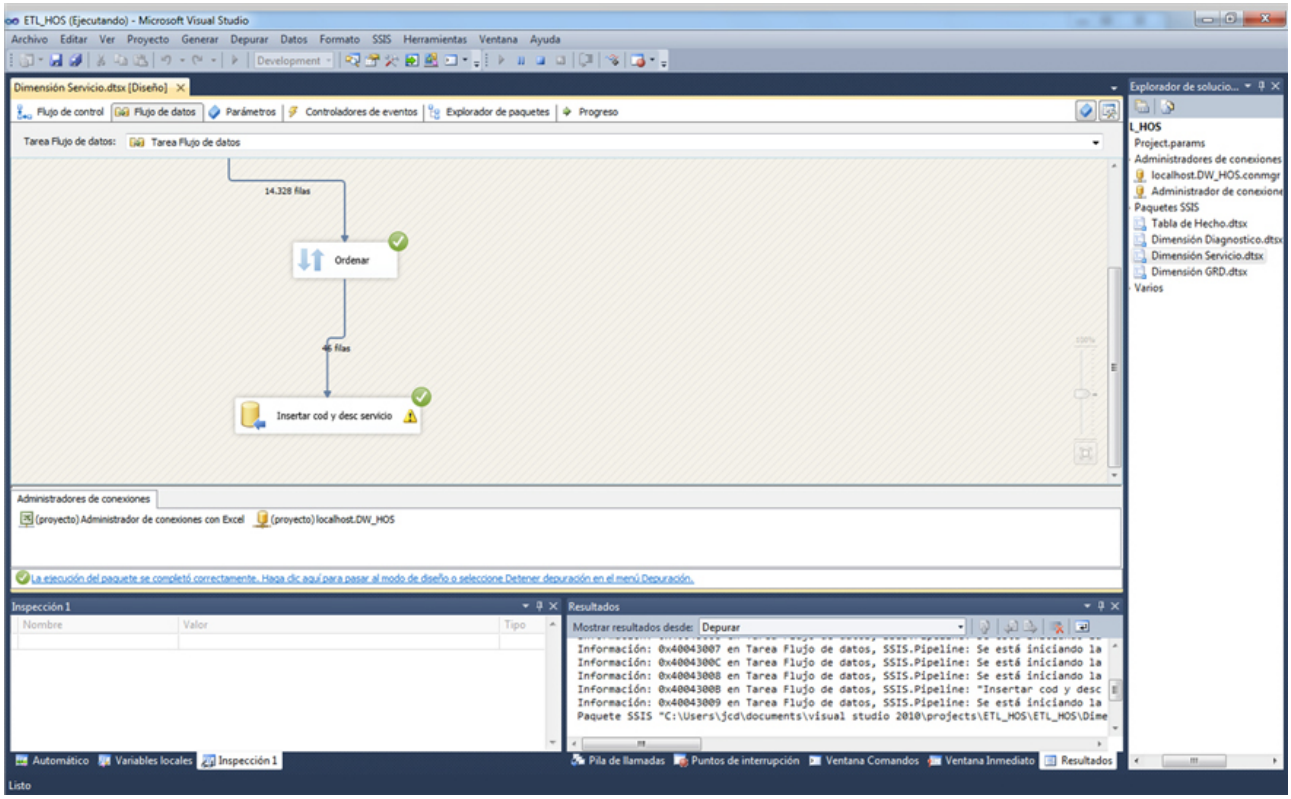
Creamos un nuevo proceso ETL denominado dimensión servicio, copiamos el paso Excel que tenemos ya configurado y quitamos todos los campos exceptuando cod_servicio y desc_servicio. Añadimos el paso de ordenar y quitamos duplicados. Finalmente, añadimos el paso de insertar los registros.



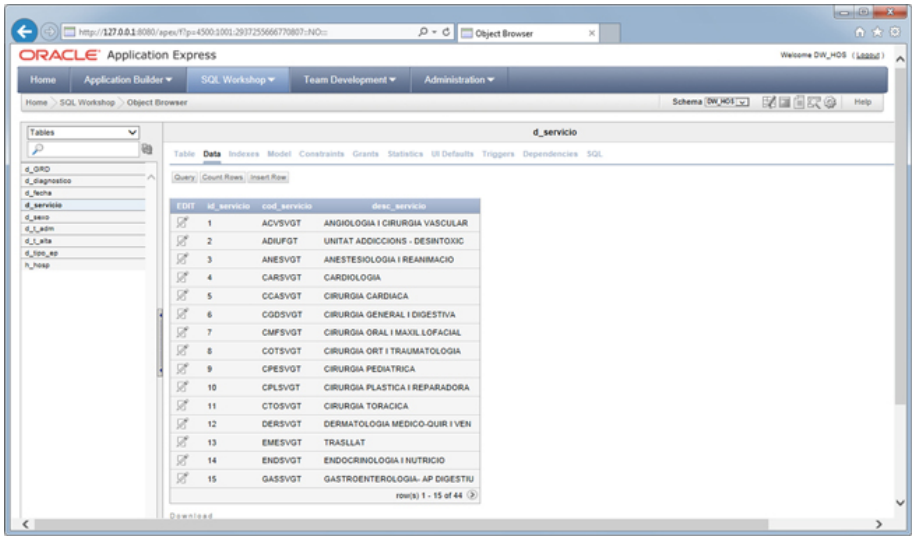
Después, añadimos el paso de insertar los registros.



Ejecutamos.

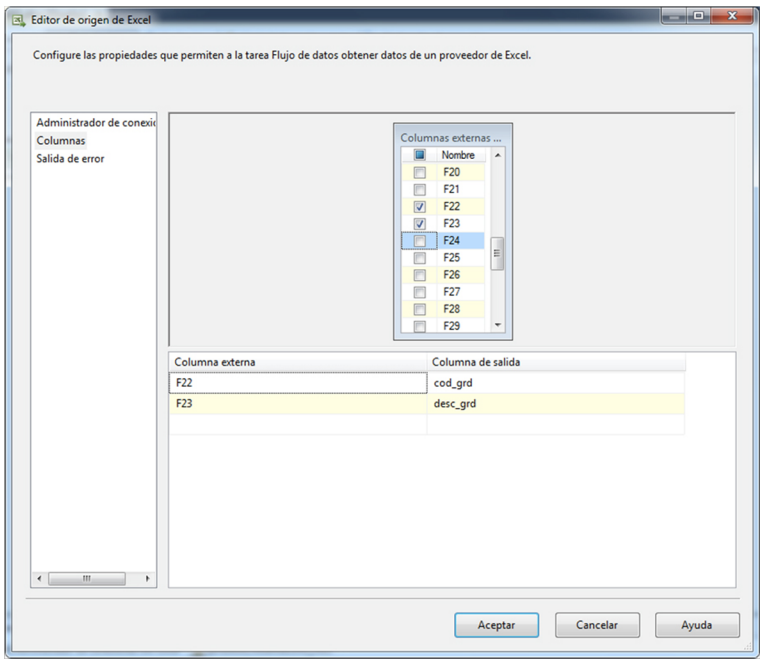


Y el resultado de la inserción son 44 registros.

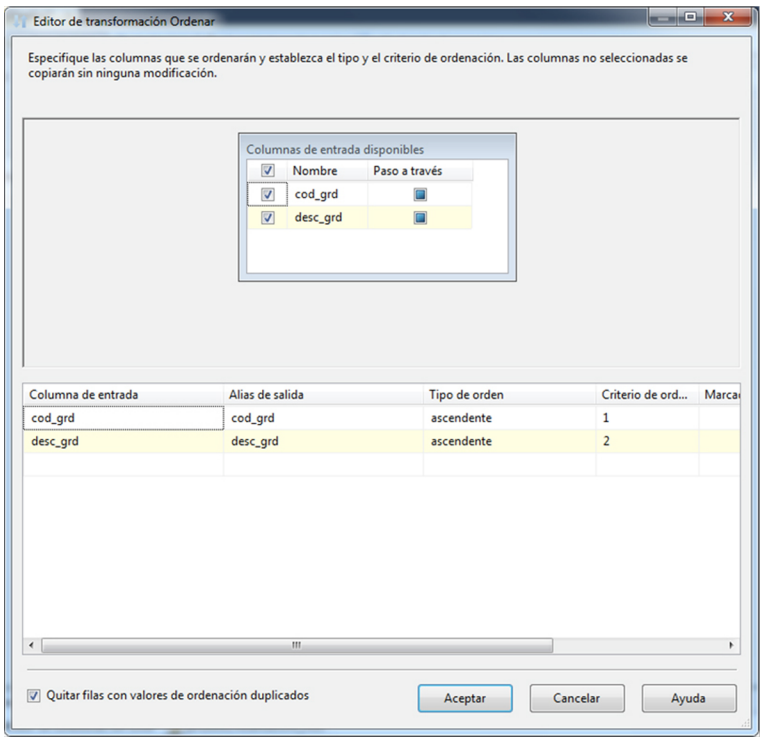


4.2.4. Dimensión GRD

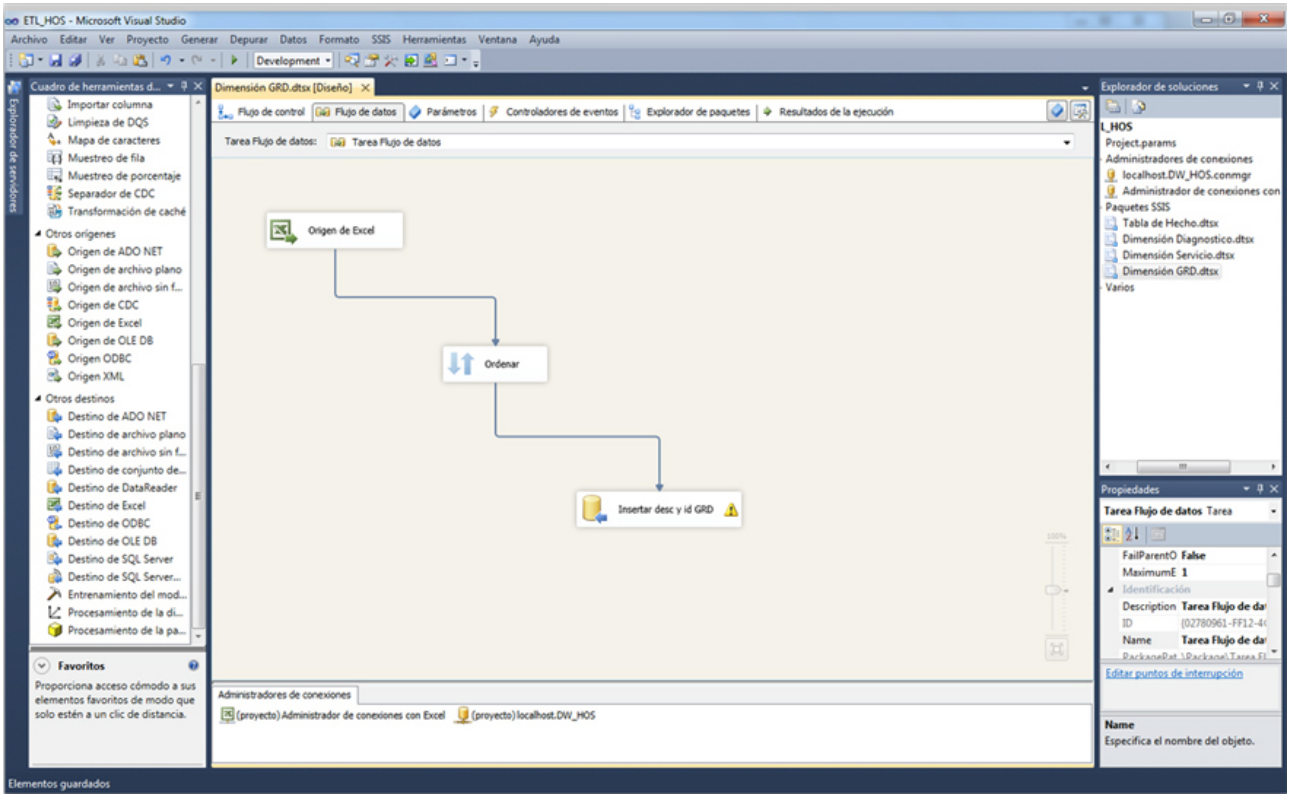
Creamos un nuevo proceso ETL llamado dimensión GRD, copiamos el paso Excel que tenemos ya configurado y quitamos todos los campos exceptuando cod_grd y desc_grd.



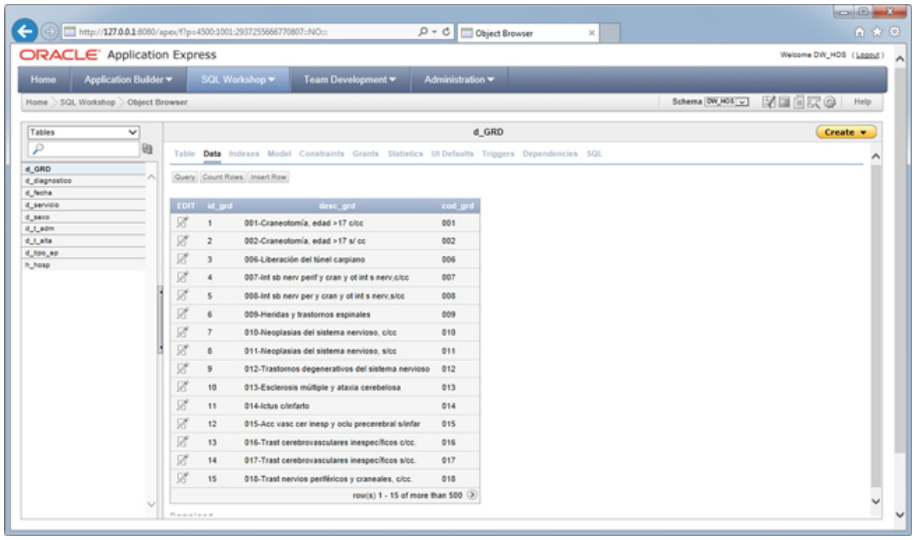
Añadimos el paso de ordenar y quitar duplicados.



Añadimos el paso de insertar registros en la base de datos y, tras configurarlo, nos quedará lo siguiente.



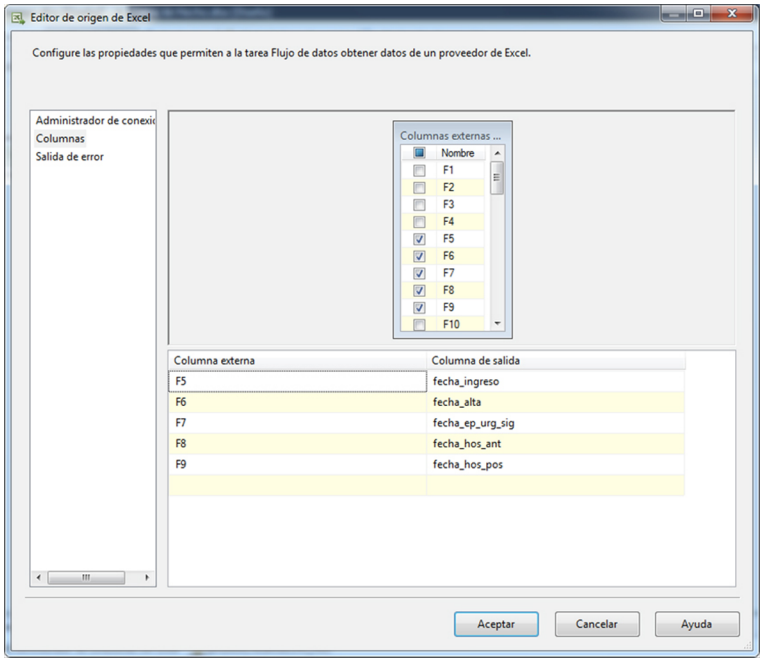
Al ejecutar, habremos insertado los valores para GRD. Hemos insertado 565 registros.



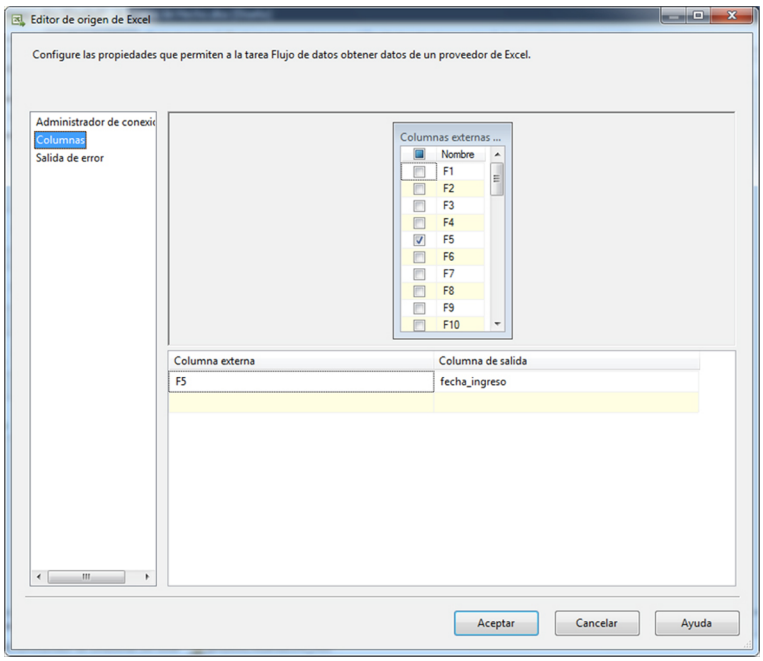
4.2.5. Dimensión fecha

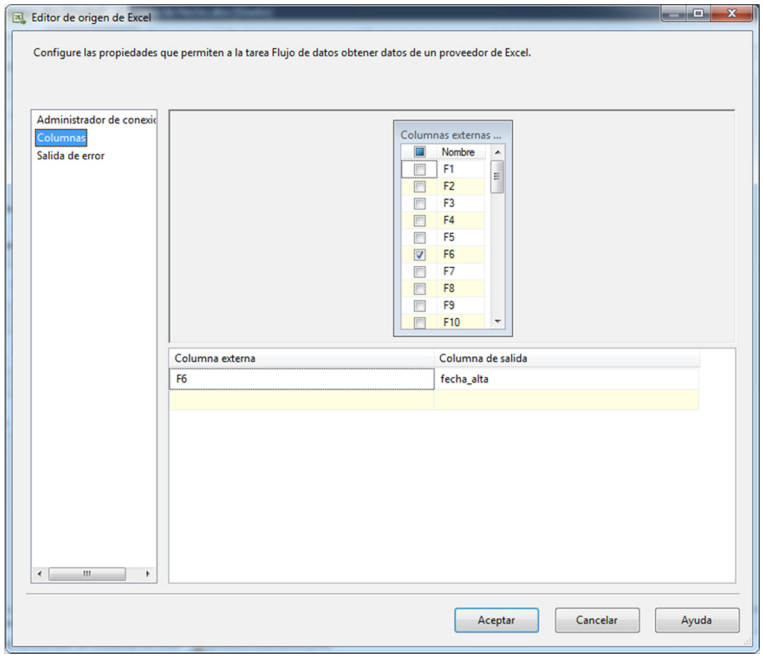
Finalmente, podemos centrarnos en la última dimensión: la dimensión tiempo. Creamos un nuevo proceso ETL denominado dimensión tiempo. Esta dimensión debe contener todas las referencias de tiempo que están en el fichero Excel.

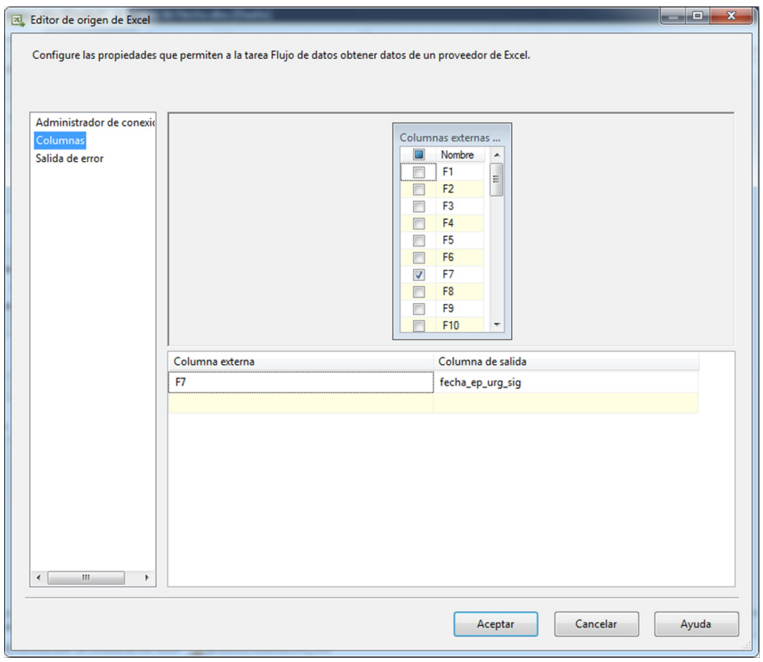
La estrategia que vamos a seguir para hacer este proceso ETL es la siguiente. Vamos a extraer cada una de las fechas y vamos a crear una combinación de las mismas. Inicialmente, del fichero Excel podemos extraer fecha_ingreso, fecha_alta, fecha_ep_urg_sig, fecha_hos_ant y fecha_hos_pos.

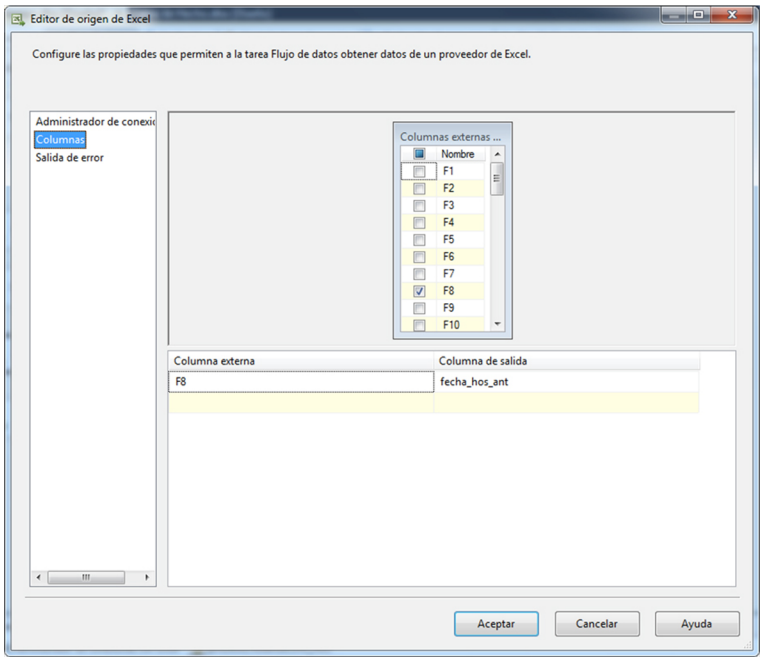


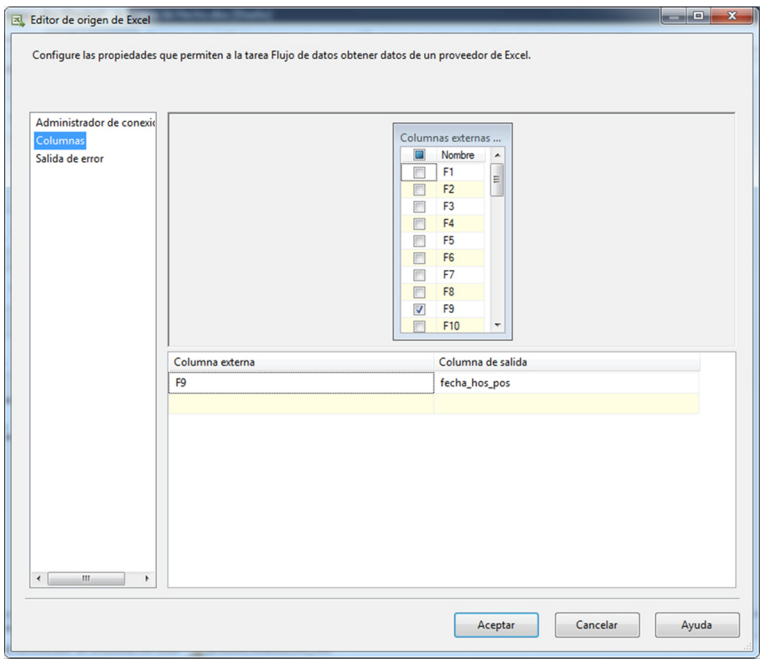
Lo que se traduce en:


Editor de origen de Excel
Configure las propiedades que permiten a la tarea Flujo de datos obtener datos de un proveedor de Excel.
Administrador de conexi
Columnas
Salida de error
Columnas externas ...
Nombre
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
Columna externa
Columna de salida
F5
fecha_ingreso
Aceptar
Cancelar
Ayuda

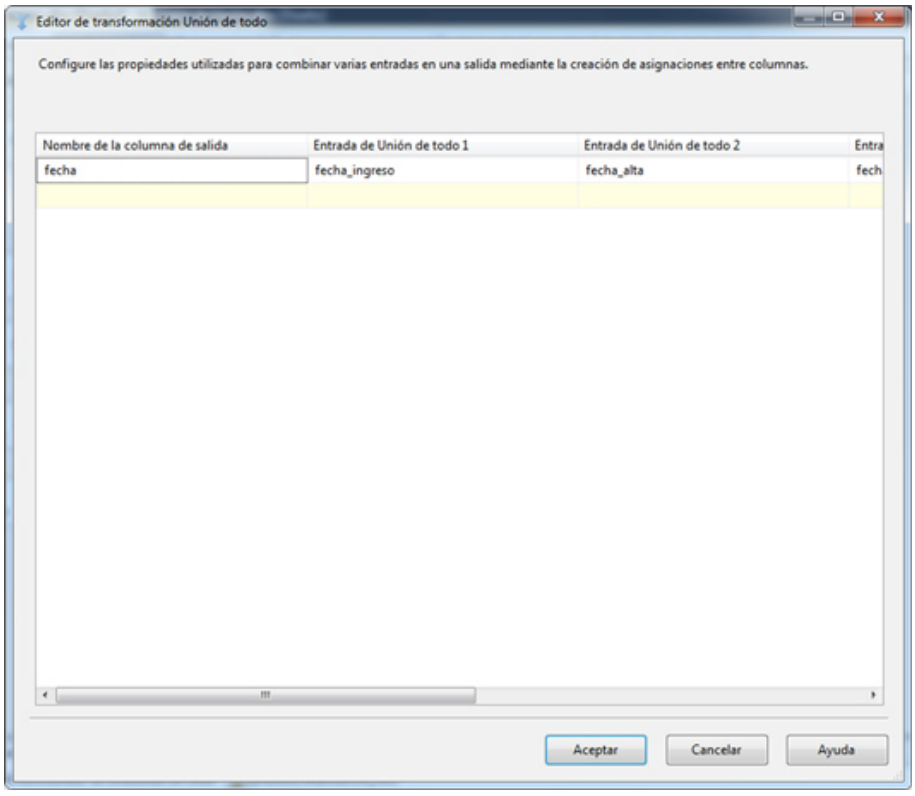

Editor de origen de Excel
Configure las propiedades que permiten a la tarea Flujo de datos obtener datos de un proveedor de Excel.
Administrador de conexi
Columnas
Salida de error
Columnas externas ...
Nombre
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
Columna externa
Columna de salida
F6
fecha_alta
Aceptar
Cancelar
Ayuda


Editor de origen de Excel
Configure las propiedades que permiten a la tarea Flujo de datos obtener datos de un proveedor de Excel.
Administrador de conexi
Columnas
Salida de error
Columnas externas ...
Nombre
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
Columna externa
Columna de salida
F7
fecha_ep_urg_sig
Aceptar
Cancelar
Ayuda

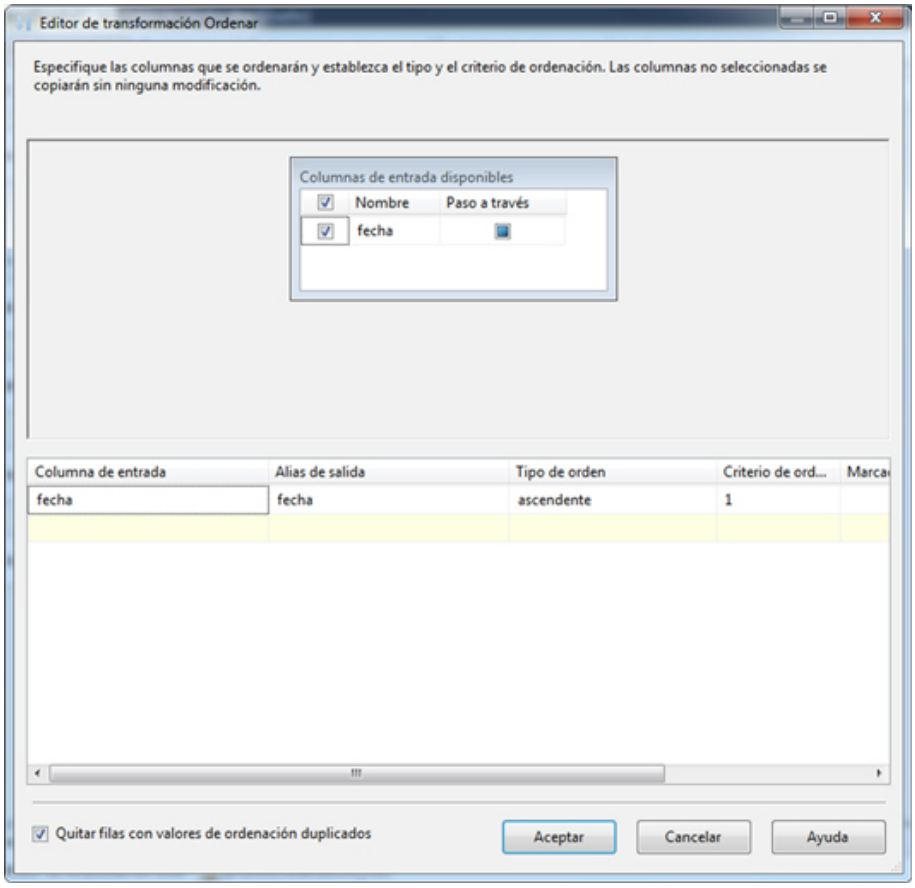

Editor de origen de Excel
Configure las propiedades que permiten a la tarea Flujo de datos obtener datos de un proveedor de Excel.
Administrador de conexi
Columnas
Salida de error
Columnas externas ...
Nombre
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
Columna externa
Columna de salida
F8
fecha_hos_ant
Aceptar
Cancelar
Ayuda



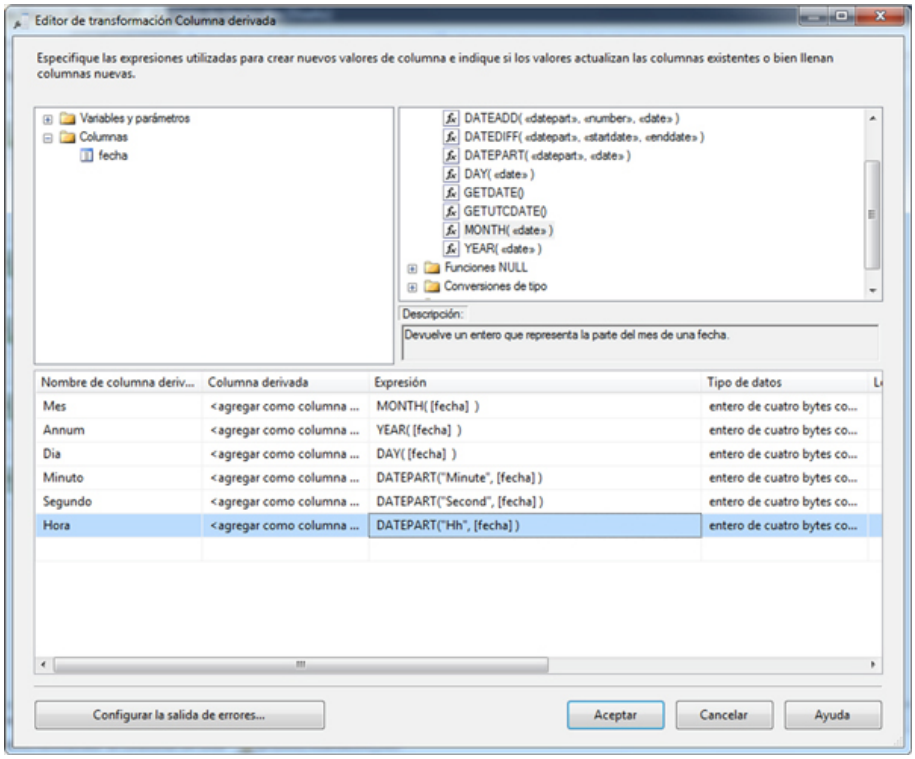
Ahora usamos un paso para unir los cinco flujos de datos bajo un mismo flujo con un único campo: fecha.



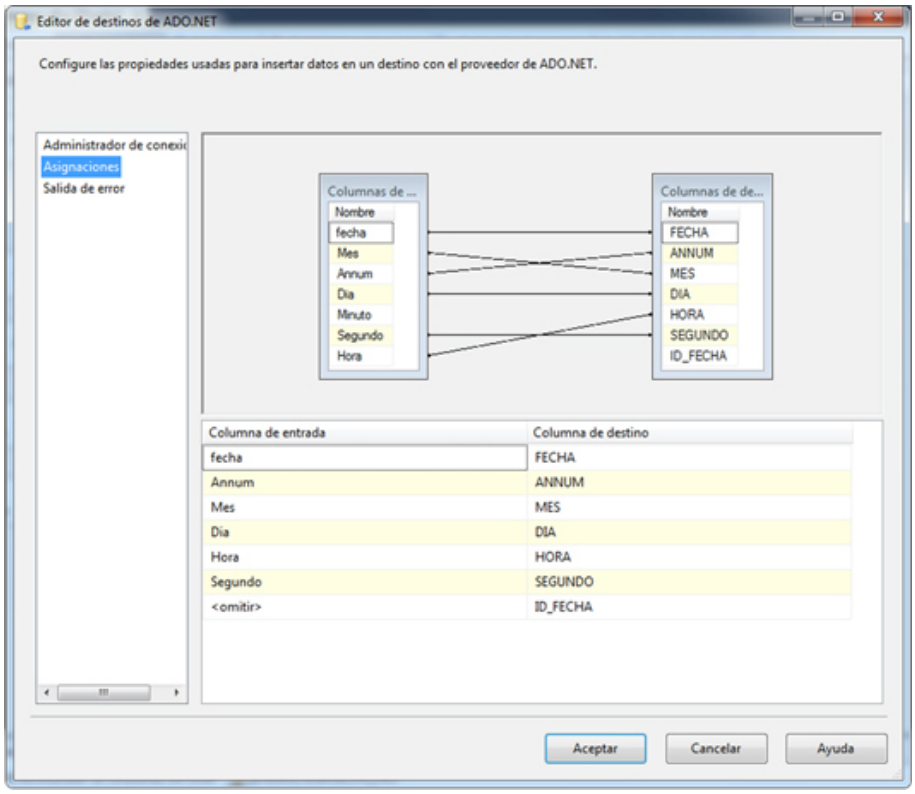
Ordenamos y quitamos duplicados.



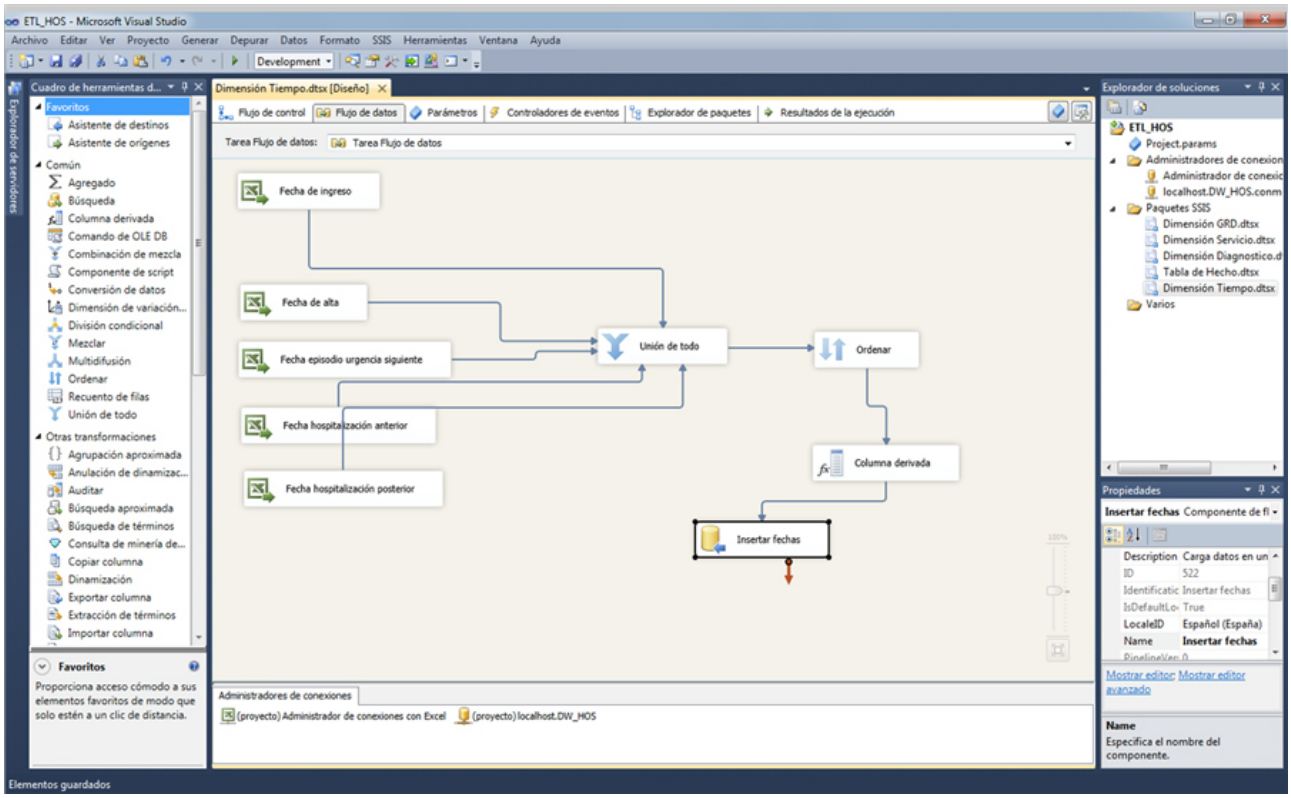
Con el paso columna derivada generamos las columnas año, mes, hora, minuto y segundo a partir de la fecha.



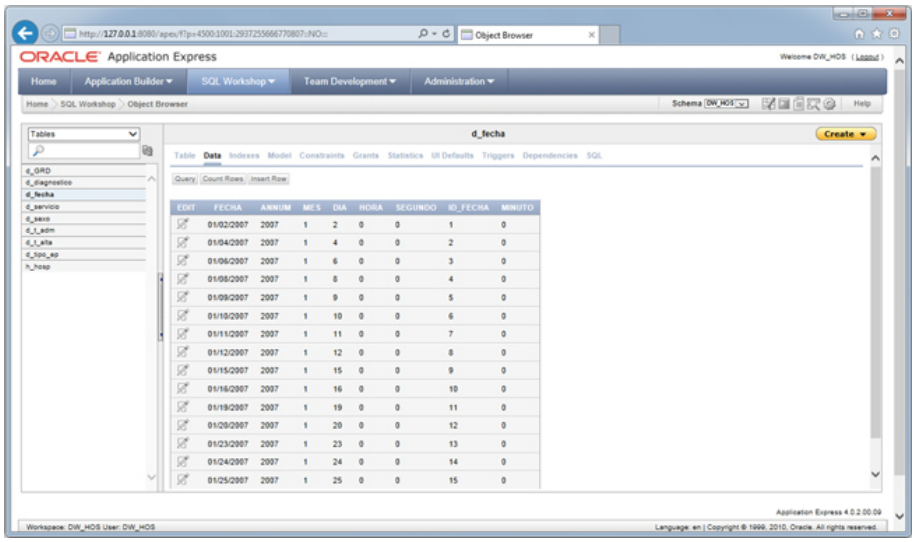
Finalmente, añadimos un paso para insertar los registros.



El proceso ETL resultante final es:



Por último, como en casos anteriores insertamos los registros. El resultado es la inserción de 30.134 registros.

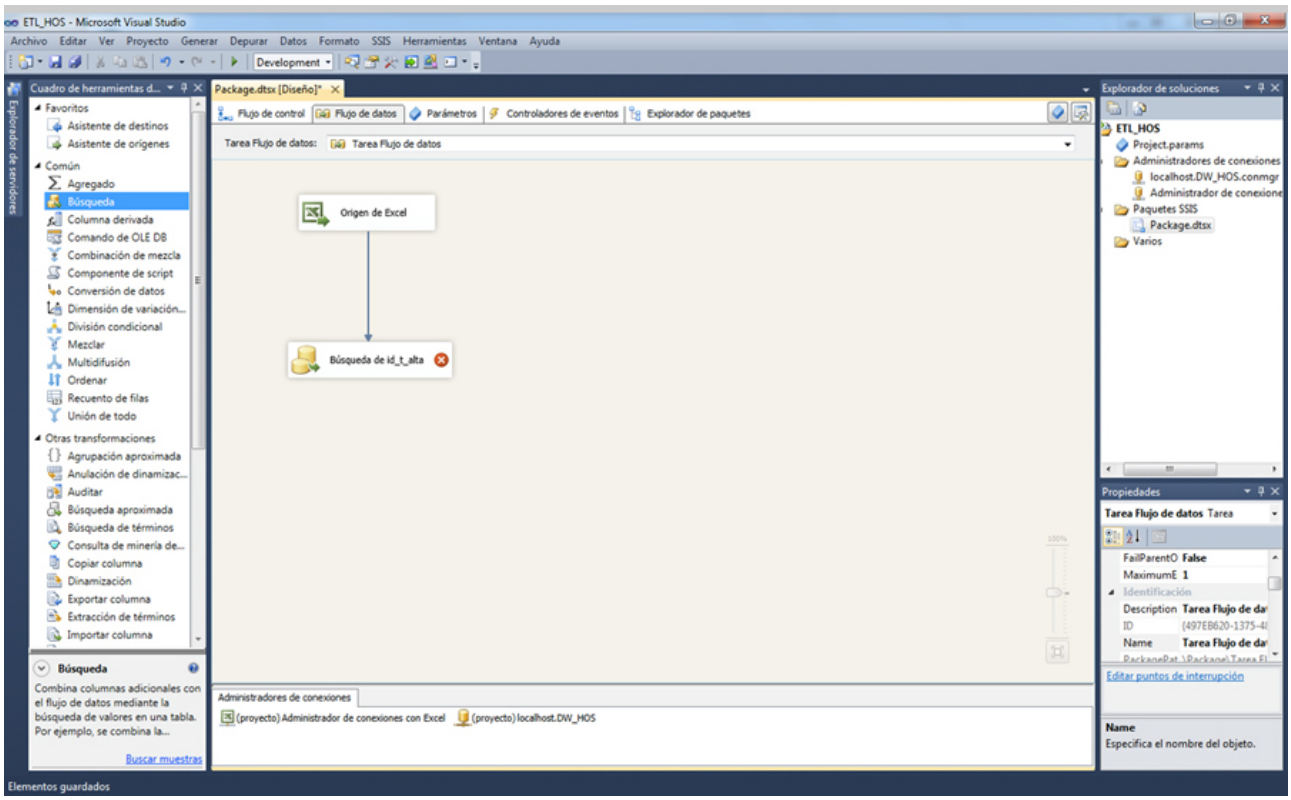


Cabe comentar que la dimensión temporal puede tener otro diseño. El diseño propuesto va a crecer en número de registros rápidamente. Para optimizar esta dimensión, sería necesario diseñarla en forma de estrella separando año, mes, día, hora, minuto y segundo.

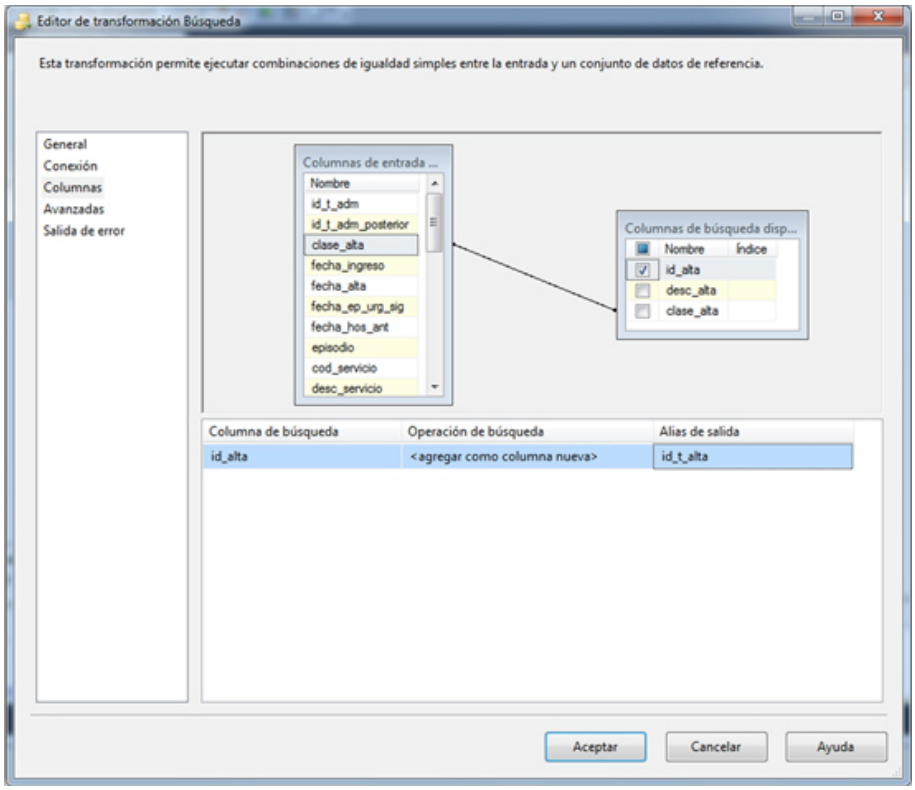
4.2.6. Tabla de hechos hospitalización

Finalmente, nos queda crear el proceso ETL para la tabla de hechos. Primero, usamos el fichero Excel. Esta vez no omitimos campos del fichero, sino que recuperamos una tras otra las claves primarias de las dimensiones por cada hecho.

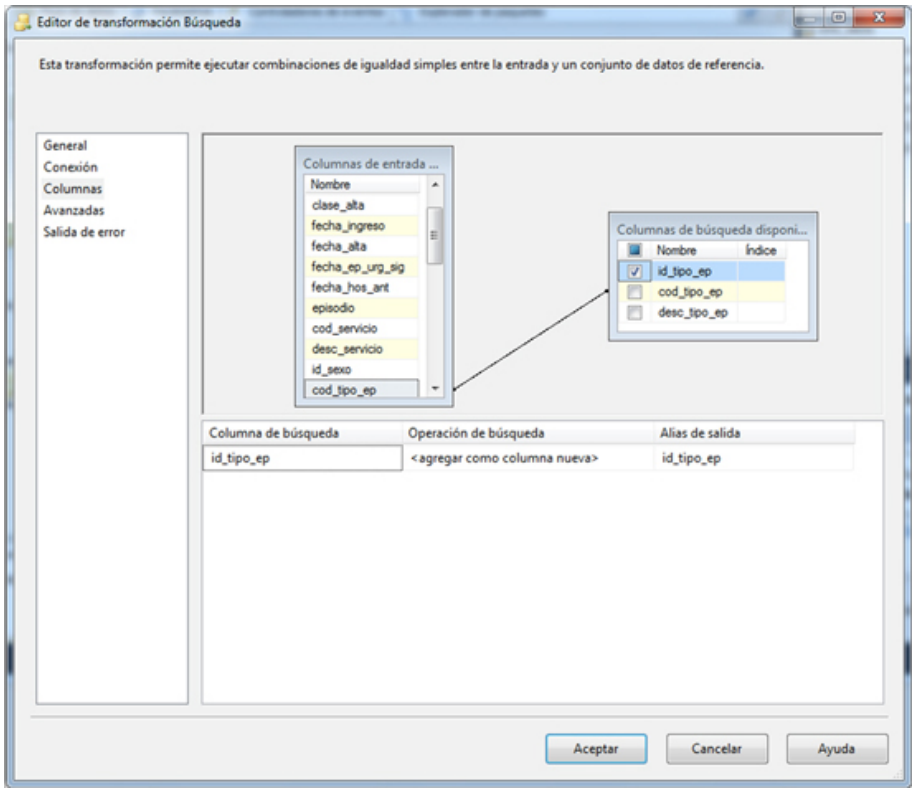
Recuperamos id_t_alta a partir de clase_alta. Para esto, haremos una búsqueda en la dimensión d_t_alta y recuperaremos la clave primera.



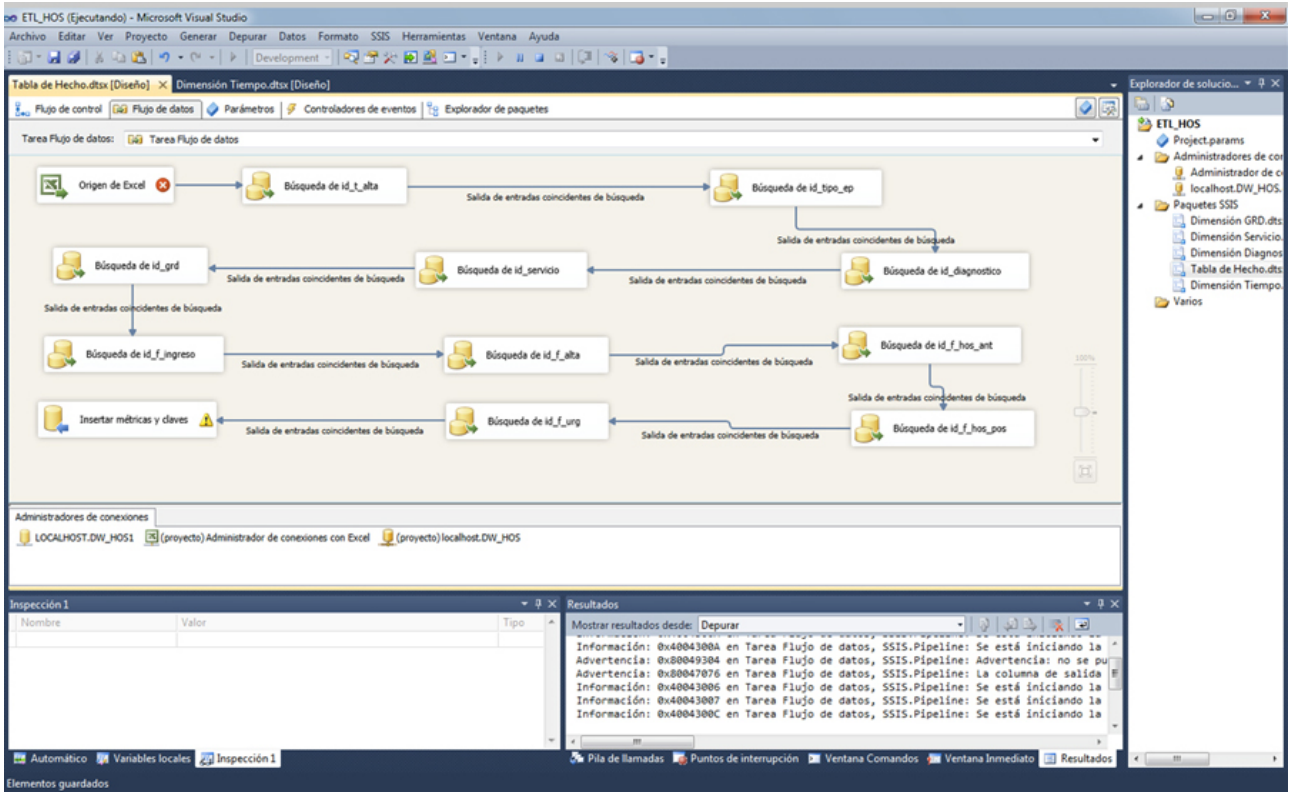
Debemos parametrizar el paso y elegir la conexión de base de datos que hay que utilizar, la tabla de la que vamos a extraer información, los campos que sirven para hacer la búsqueda y el campo que recuperamos.



Para recuperar id_tipo_ep a partir de cod_tipo_ep, hacemos lo mismo pero esta vez usando la dimensión d_tipo_ep.

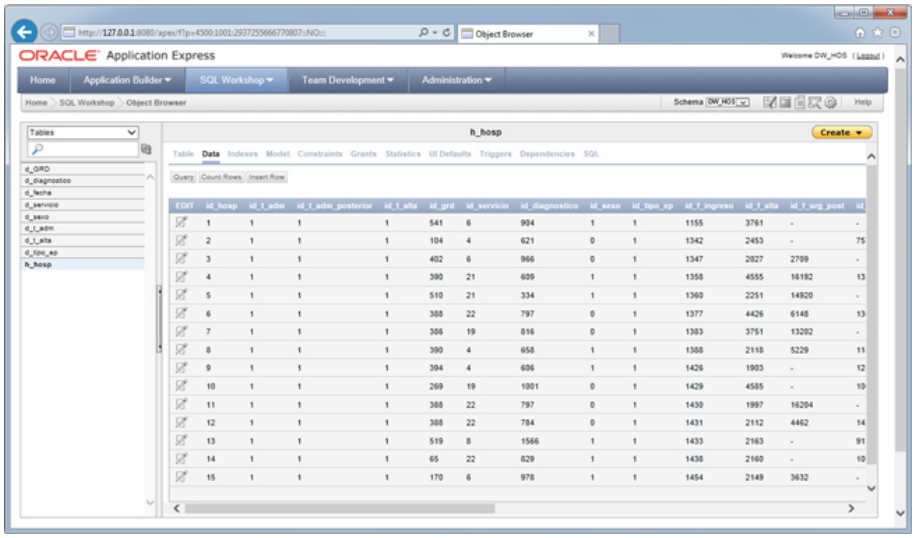


El proceso, repetido tantas veces como dimensiones, se culmina con el paso final de inserción en la base de datos. Esto nos da la transformación siguiente.



Al ejecutarla, habremos completado los procesos ETL para alimentar el almacén de datos.

El resultado del ETL es la inserción de 14.326 registros:



4.2.7. Consideraciones finales

Tras el diseño de los procesos de carga y la ejecución de los mismos, podemos conocer el tamaño del almacén de datos. Esta información va a permitir a futuro optimizar la factoría de información. Por ejemplo, teniendo en cuenta el tamaño de los registros para este ejemplo no se propone una vista por años, pero quizá sería conveniente en el ejemplo real.

Una vez tenemos la información cargada en nuestro almacén de datos, ya podemos pasar a la fase de explotación.

5. Explotación de datos

5.1. Requerimientos y usuarios

Una vez efectuada la carga de datos en nuestro *data warehouse*, es el momento de crear un modelo OLAP. Antes de ponernos a crear el modelo OLAP, debemos identificar qué usuarios son los susceptibles de usar análisis multidimensional, teniendo en cuenta que este tipo de usuarios son avanzados y, por lo tanto, buscan analizar ellos mismos la información.

Como se apuntó en el análisis de requerimientos, los usuarios del sistema son la dirección clínica y el propio centro. Dentro de la dirección clínica y el propio centro, los potenciales tipos de usuario que pueden usar el modelo OLAP son distintos.

- Responsable del área de hospitalización: este usuario busca comprender la evolución del área de hospitalización y tener claro qué servicios son los más demandados, para planificar de manera adecuada los servicios que se ofrecen. El usuario consultará la información directamente o por medio de la figura de un analista de negocio.

- Dirección clínica: mediante la figura de un analista de negocio, se busca entender cómo ha evolucionado el área de hospitalización dentro del conjunto de áreas del hospital, y cómo lo que sucede en la misma afecta al comportamiento global del hospital.

Aparte, existen otros usuarios que se pueden beneficiar del *data warehouse*.

- Departamento financiero: cada tipo de estadía tiene asociado un coste para el hospital. Pese a tener un fin social, el hospital debe controlar correctamente los costes asociados por paciente para garantizar su sostenibilidad. Por tanto, el analista financiero/contable deberá ser capaz de cruzar esta información con los costes, para identificar si se ha seguido el presupuesto planeado, calcular los costes reales y hacer planificaciones futuras.

Para la creación del modelo OLAP vamos a hacer uso de Visual Studio 2010, que nos va a permitir generar de manera sencilla una estructura MOLAP basada en las tablas en Oracle que hemos diseñado con anterioridad.

Una pregunta natural que surge es la siguiente: ¿qué sentido tiene crear un modelo OLAP? Crear un modelo OLAP significa crear un cubo que represente el *data warehouse* o, en general, un subconjunto del mismo. La importancia de crear este elemento de análisis es múltiple:

- Evitar el acceso de los usuarios a información más allá de su alcance de negocio.

- Precalcular todas las posibles consultas mutidimensionales (reduciendo problemas de rendimiento).

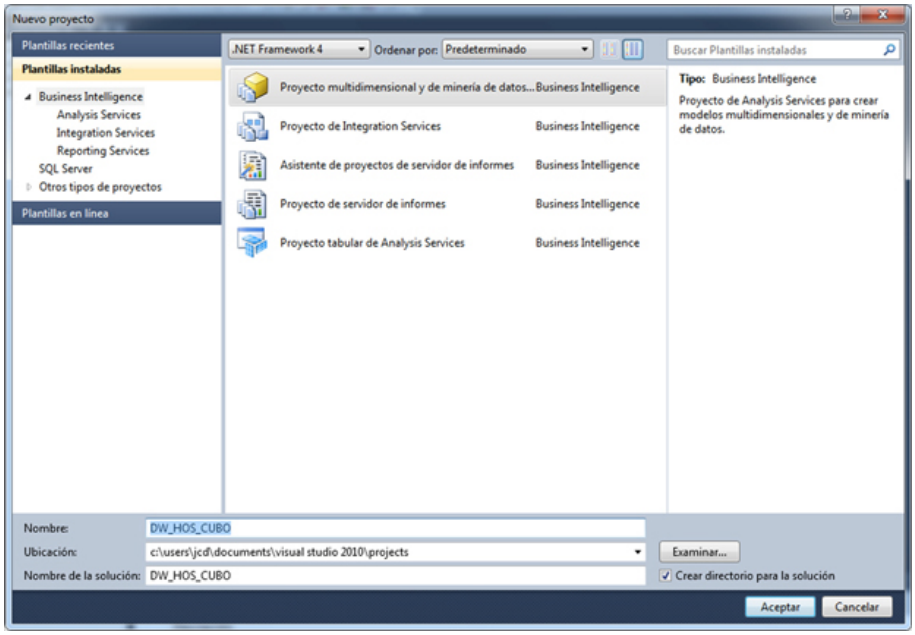
- Proporcionar una fuente de datos "transportable" y *offline* para los usuarios. Por ejemplo, un usuario comercial podría llevarse una copia del cubo en su ordenador en el caso de necesitarla, y acceder a la misma mediante Excel.

- Proporcionar una herramienta de análisis no predefinida a los usuarios que necesitan una herramienta más allá de los informes.

5.2. Modelo OLAP

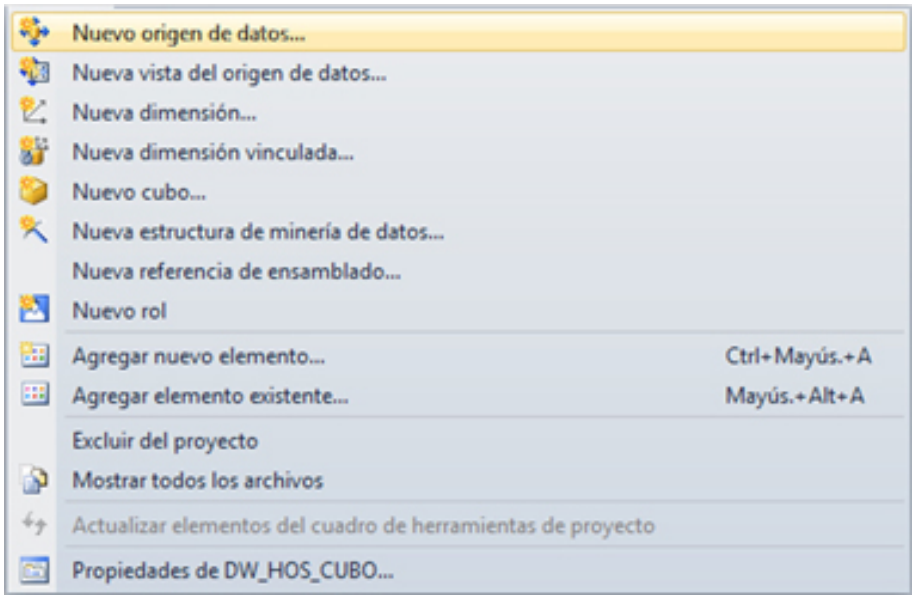
1) Creando el proyecto

Iniciamos Visual Studio y seleccionamos un nuevo proyecto. El tipo de proyecto que nos interesa esta vez es el "proyecto multidimensional y de minería de datos", que permite tanto crear cubos como proyectos de minería de datos. Por ejemplo, le damos a este proyecto el nombre de DW_HOS_CUBO.

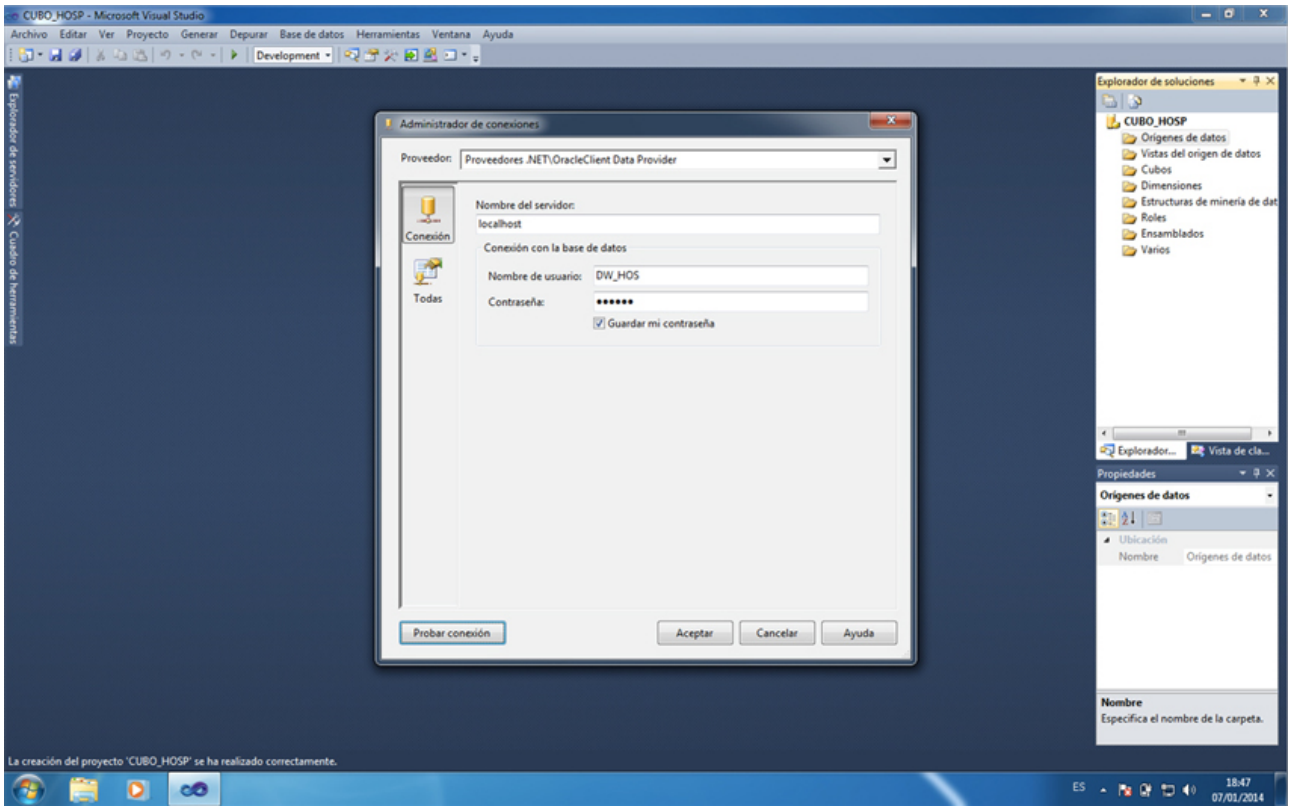


2) Estableciendo el origen de datos

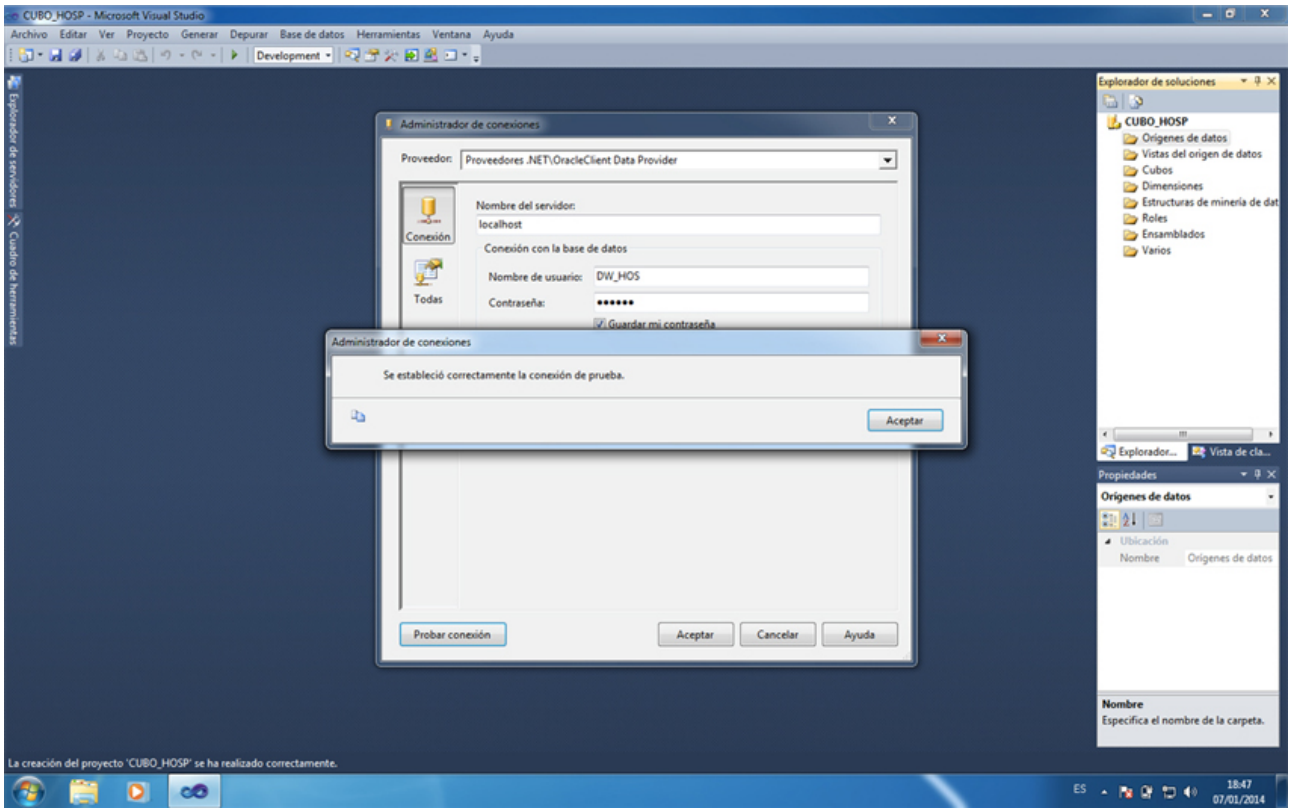
Una vez creado el proyecto, tenemos su estructura en el lateral. Esta estructura está vacía y debemos completarla para crear la estructura MOLAP. El primer paso consiste en definir el origen de los datos. Para esto, o bien hacemos clic con el botón derecho del ratón en las fuentes de origen y seleccionamos Nuevo origen de datos, o bien encontramos esta opción desplegando el menú Proyecto.



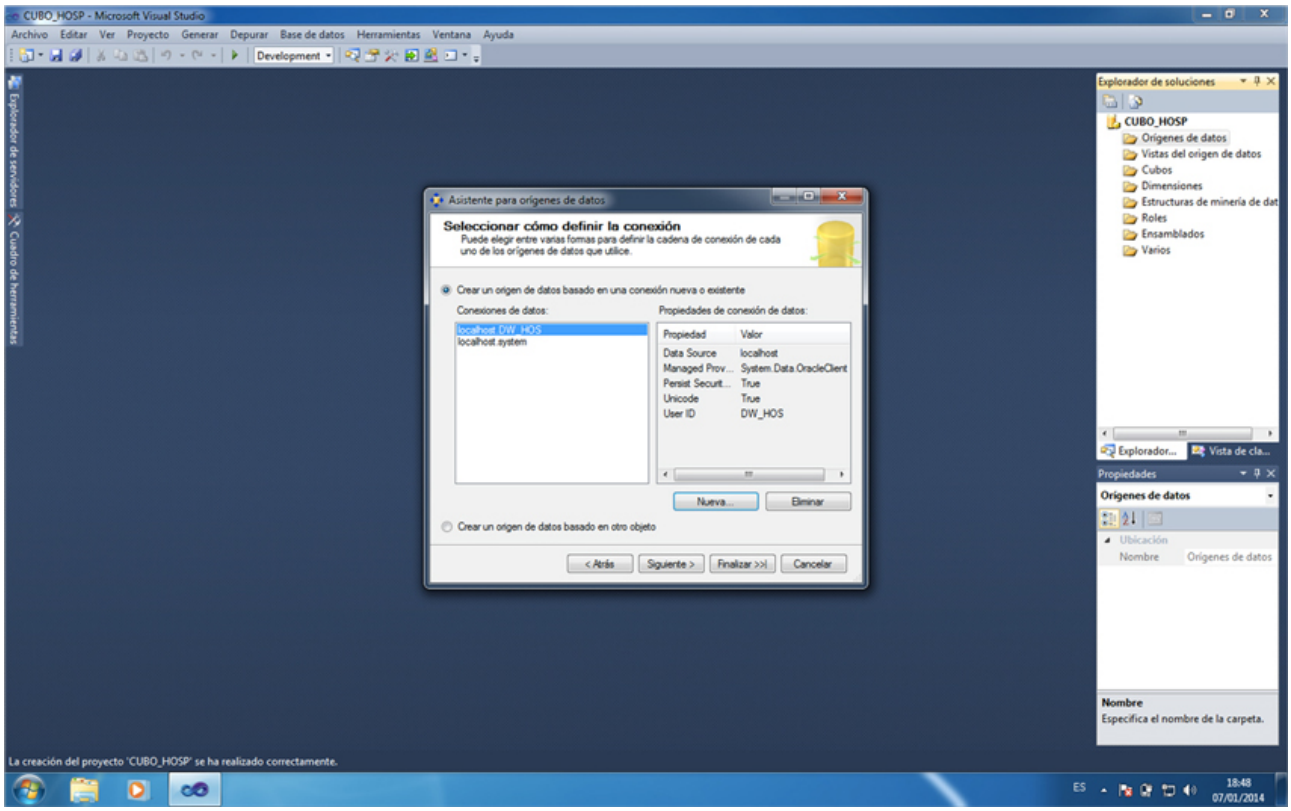
Nos aparecerá un menú para definir la conexión a base de datos. En este menú, seleccionamos Proveedor.NET\OracleClient Data Provider e indicamos la identificación y contraseña del usuario de la base de datos creada. En nuestro caso, el usuario y la contraseña tienen el mismo valor DW_HOS.



Es importante validar siempre que la conexión funciona.



Tras esto, seleccionamos la conexión que hemos creado.

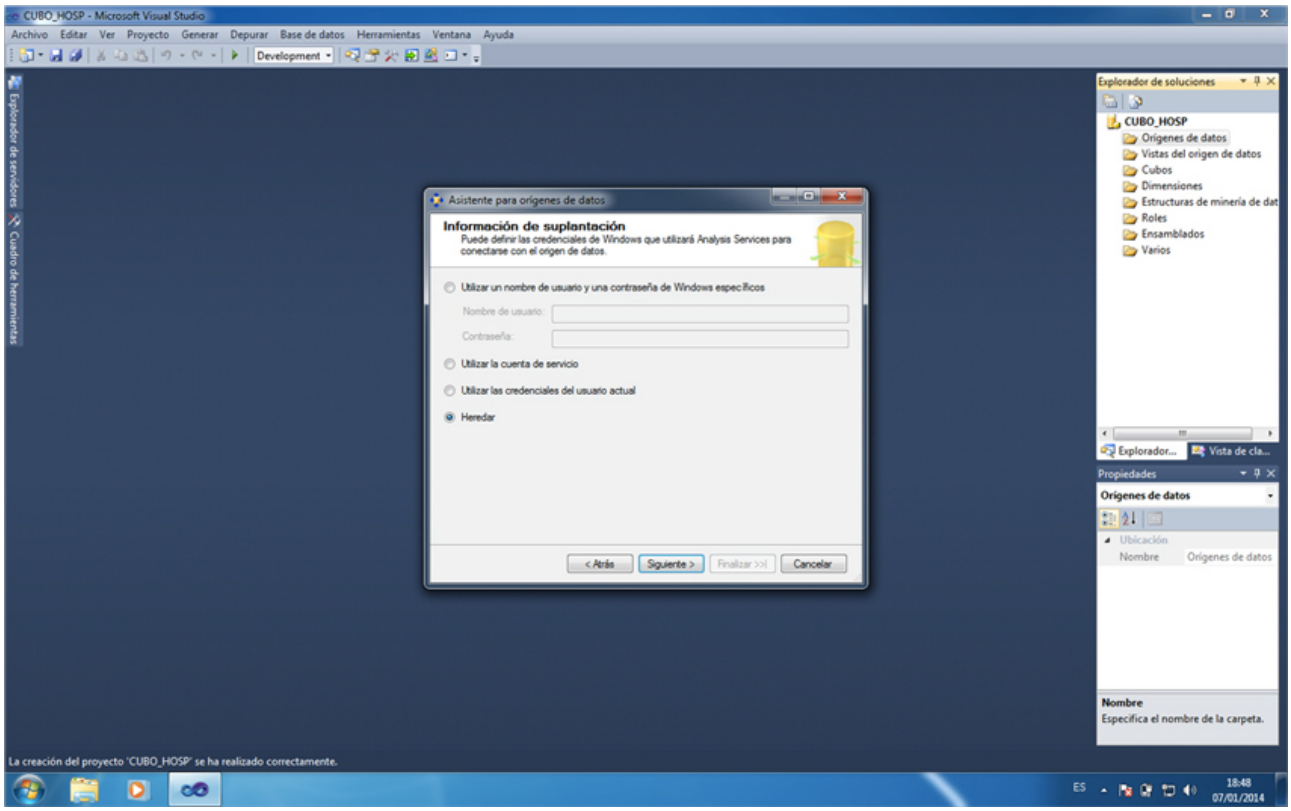


Uno de los puntos relevantes de crear elementos de análisis como un cubo es que proporcionamos accesibilidad a información a los usuarios de negocio. Por este motivo, durante el proceso de creación del cubo el asistente nos pregunta a qué usuarios les proporcionaremos acceso. En nuestro caso, no tenemos un listado de usuarios o grupo de usuarios en un LDAP[8] por lo que, para facilitar la creación, utilizaremos el usuario por defecto del sistema operativo.

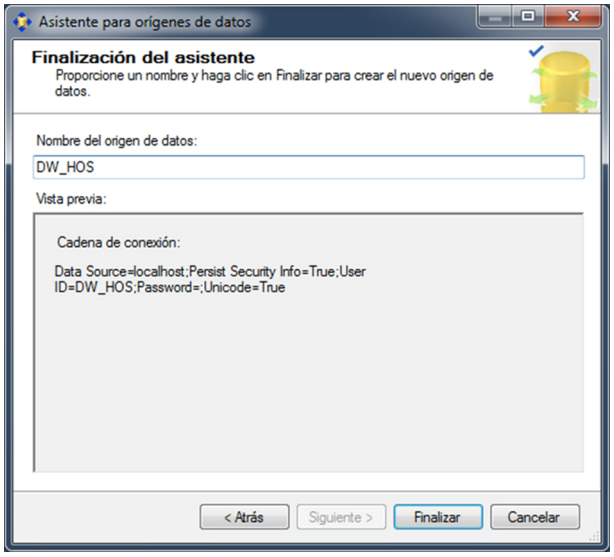
[8] *LDAP* son las siglas de *lightweight directory access protocol.*

Por lo tanto, indicamos que heredamos el usuario del sistema para el cubo[9]. Esta selección de usuarios siempre se puede modificar *a posteriori*.

[9] En el caso de desplegar este cubo en producción, es necesario definir qué usuarios tienen acceso a la información contenida en el cubo.



Finalmente, damos un nombre al origen de datos creado.

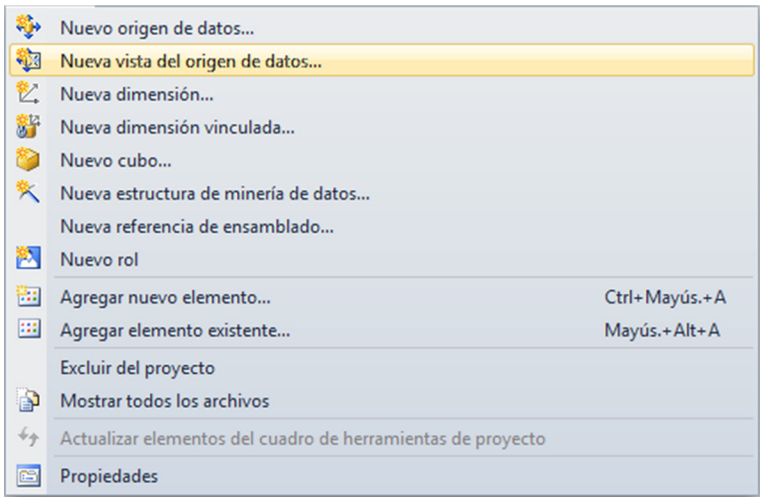


3) Creando una vista del origen de datos

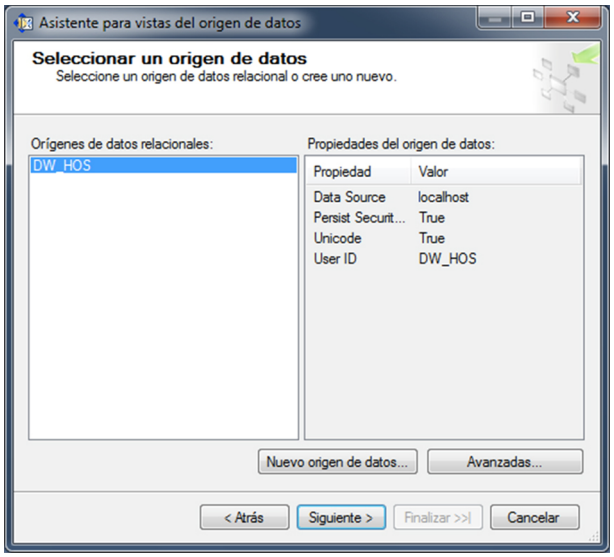
Una vez creada la fuente de origen, pasamos al siguiente paso, que consiste en crear una vista del origen de datos. Este paso suele ser muy importante en las estructuras MOLAP, dado que definen el alcance de la misma. Una vez generado el cubo, no es posible acceder a datos que no están definidos en el origen de datos. Un cubo OLAP puede tener la misma información que

la fuente de origen (en formato multidimensional), una cantidad menor o incluso una combinación de información de la fuente de origen con tablas que solo existen en el cubo.

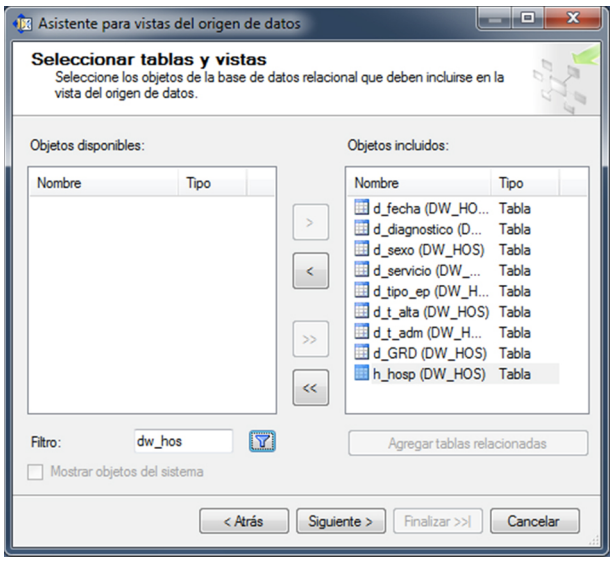
Por lo que volvemos a la estructura del proyecto y creamos una vista del origen de datos. También lo podemos hacer desplegando el menú Proyecto.



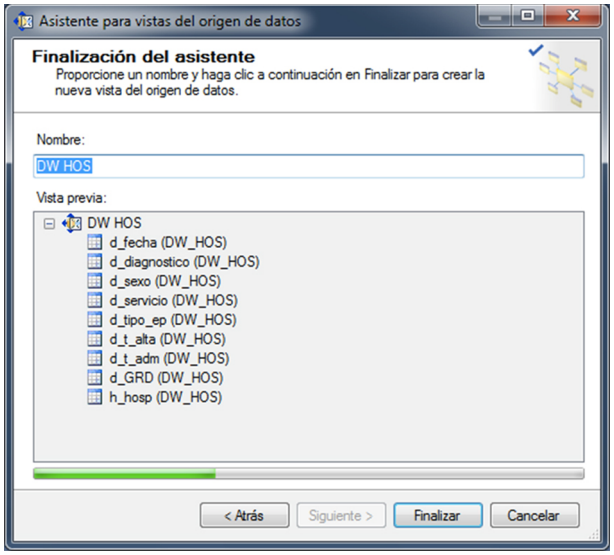
Nos aparece un asistente como en el caso anterior. Seleccionamos la fuente de origen (nos proporciona la que hemos creado).



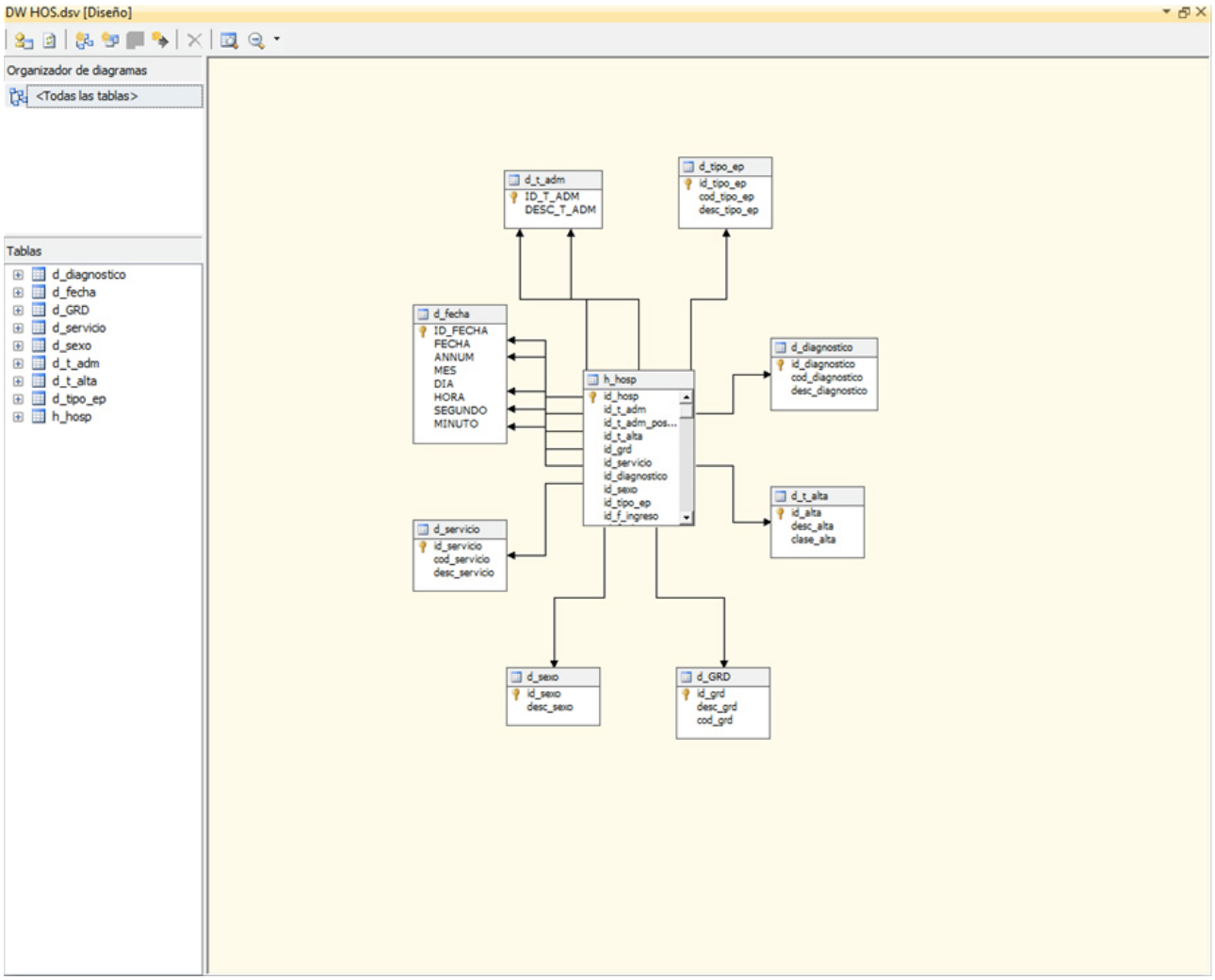
Aparecerá la lista de tablas existentes en la base de datos. Aquí, el problema reside en seleccionar aquellas que nos interesan. Afortunadamente, hay un filtro que podemos usar para seleccionarlas: DW_HOS. En nuestro caso particular, vamos a seleccionar todas las tablas que hemos creado.



Al hacer clic en Siguiente, debemos definir el nombre de la vista. Siguiendo el criterio de nombres, utilizaremos el nombre DW_HOS.

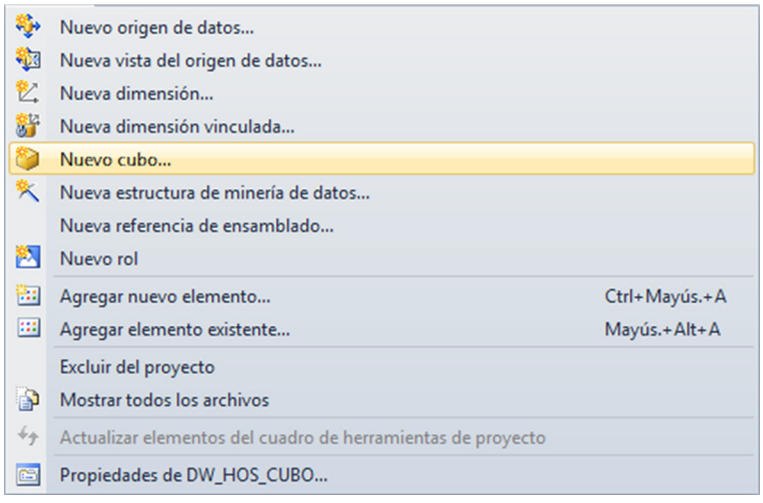


El sistema nos presenta el resultado de nuestra selección. Además, nos muestra las relaciones entre cada una de las tablas, como hemos definido por medio de las claves foráneas.

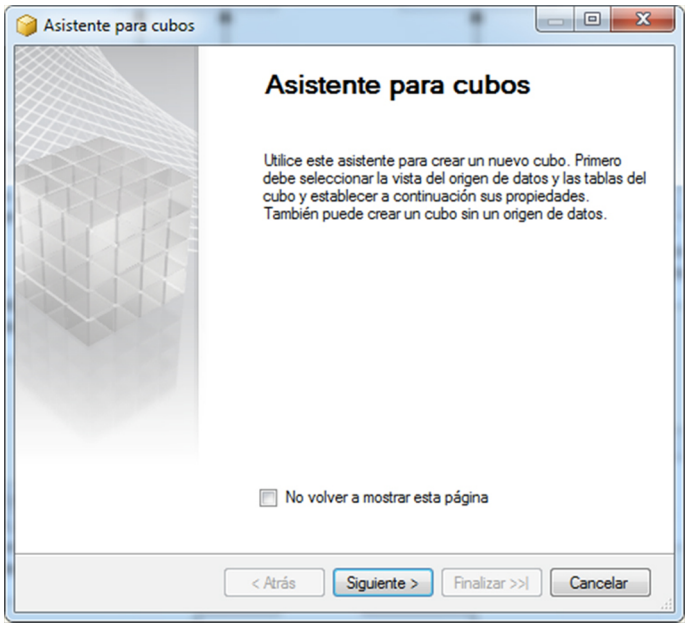


4) Creando el cubo

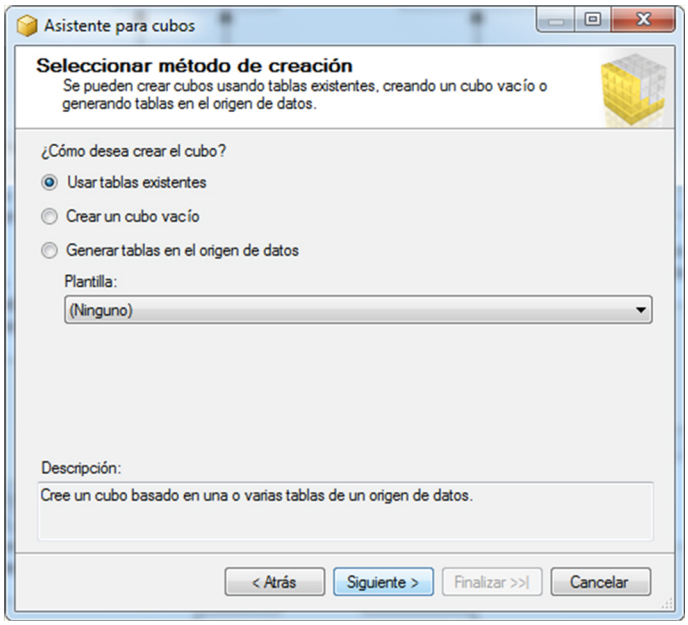
Ahora ya estamos preparados para crear el cubo. De nuevo, tenemos dos opciones: o bien mediante la estructura lateral, o de nuevo desplegando el menú de proyecto.



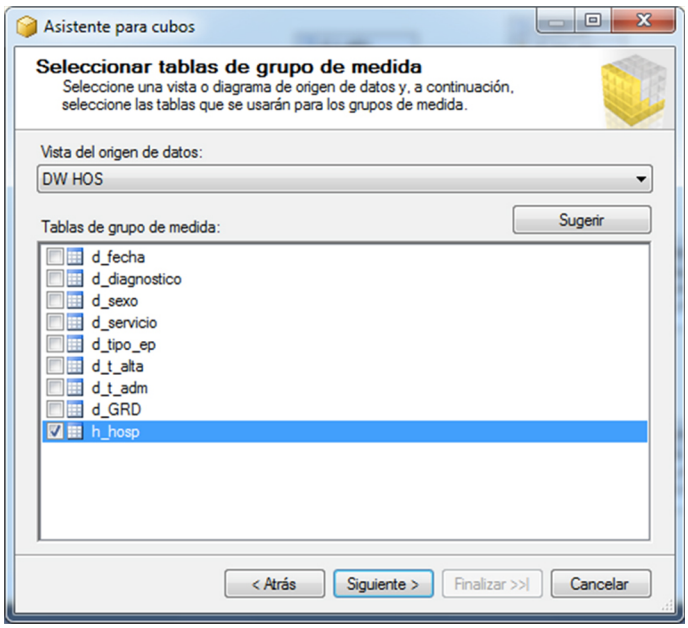
Aparecerá de nuevo un asistente.



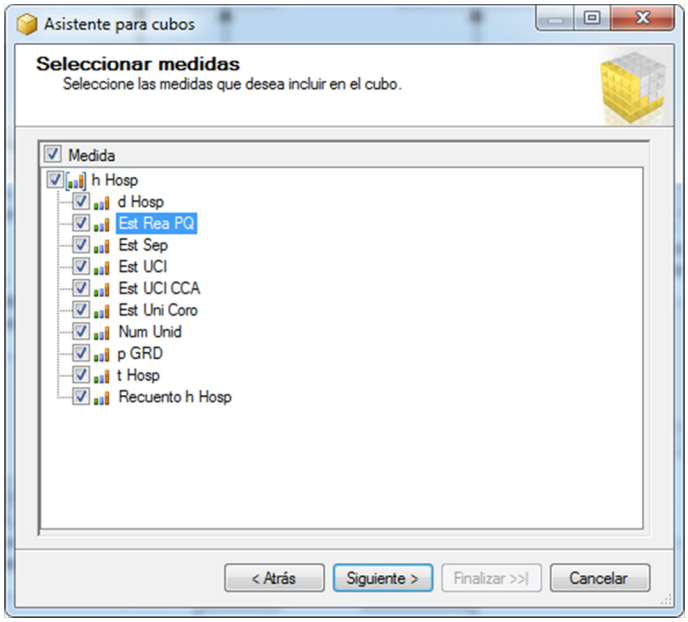
Puesto que ya hemos definido la vista y tenemos acceso a tablas, seleccionamos la opción de usar tablas existentes.



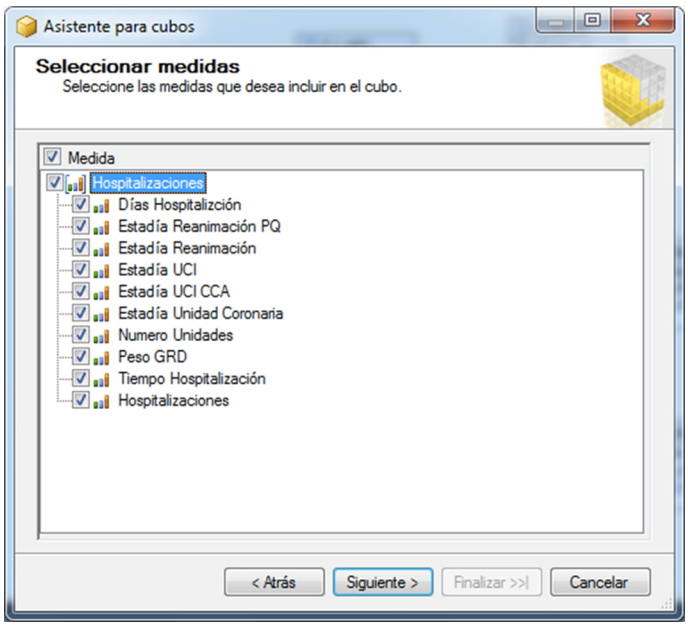
El sistema demanda que indiquemos qué tabla contiene la métrica que hay que utilizar. En nuestro caso, la tabla de hecho h_hos.



Dado que se han reconocido ya las relaciones entre tablas, el sistema propone distintas métricas de manera automática. En nuestro caso, añadiremos a su propuesta una métrica de recuento.

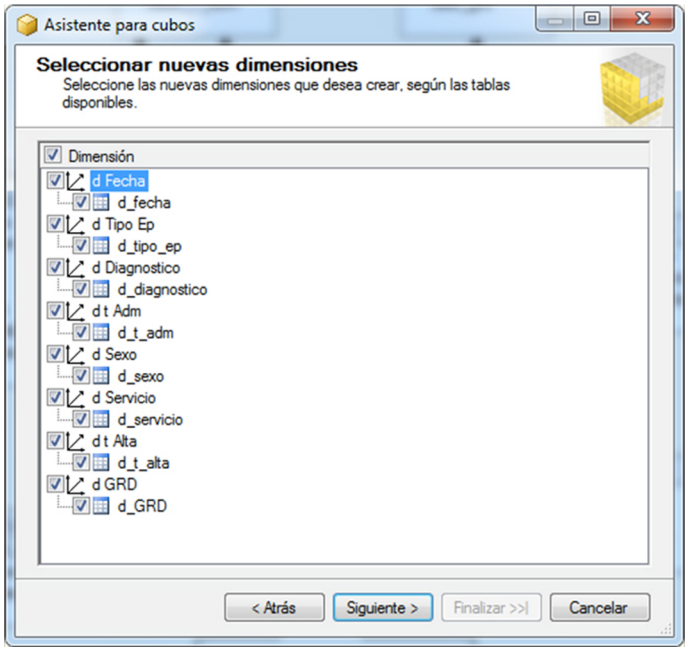


Aprovechamos que el asistente nos permite modificar el nombre de las métricas para pasar de la codificación de base de datos al lenguaje de negocio. De este modo:

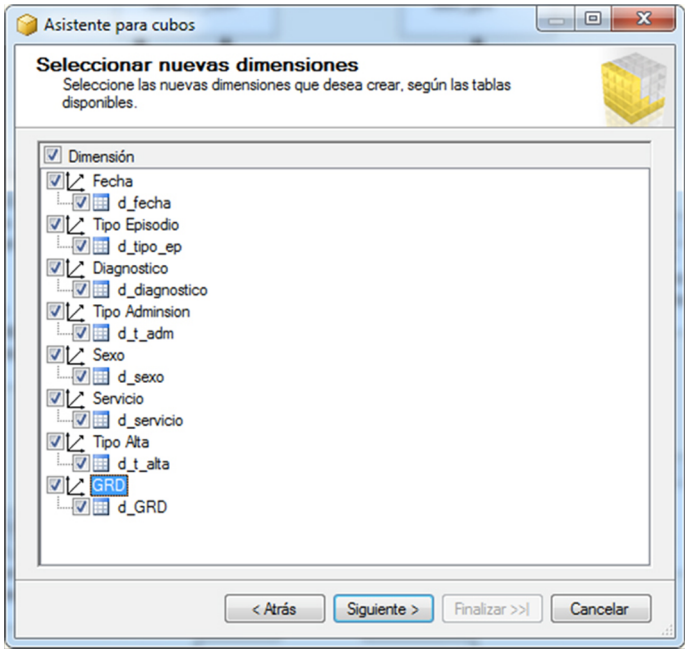


Atención: ni nhc ni episodio son métricas.

En el siguiente paso, el asistente nos propone las posibles dimensiones del cubo a partir de las claves foráneas de la tabla de hecho. En nuestro caso, aceptamos la selección propuesta. Como se aprecia, las dimensiones hacen referencia a las 8 dimensiones físicas.

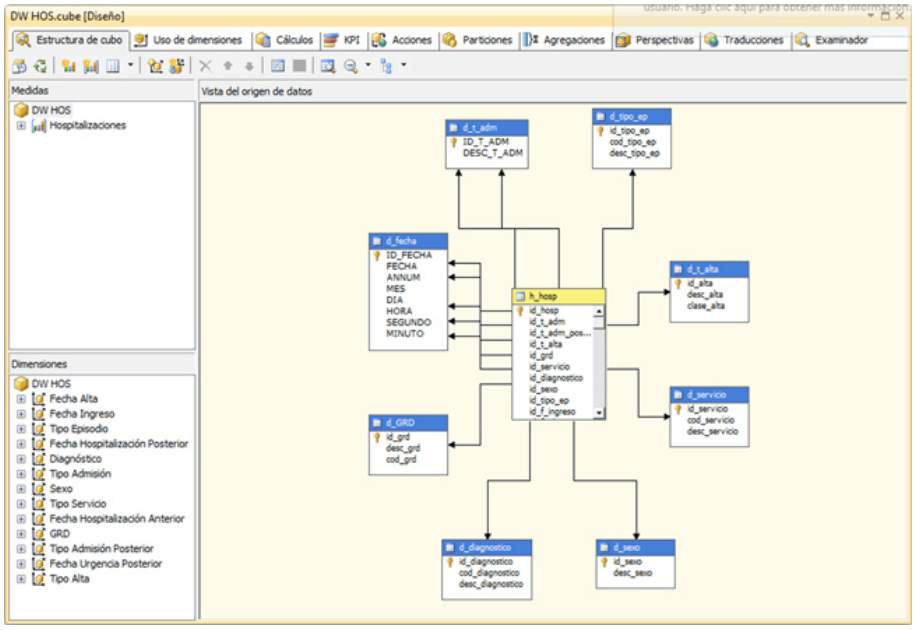


Aprovechamos para renombrar las dimensiones a términos de negocio, para facilitar su posterior uso por parte de los usuarios de negocio.

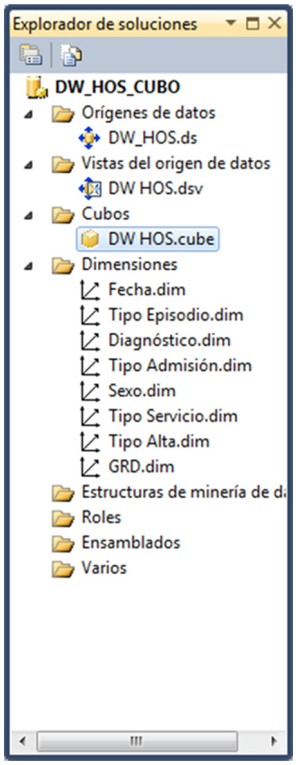


El asistente resume nuestra selección antes de terminar, y la muestra para que podamos detectar posibles errores.

El cubo resultante es el siguiente.



En el explorador, podremos ver el cubo y sus dimensiones.

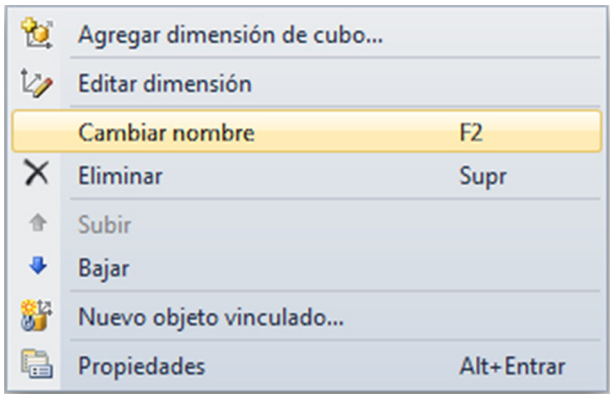


A pesar de haber creado ya el cubo mediante el asistente, no hemos terminado de trabajar con el mismo. Debemos refinar y mejorar las dimensiones. En particular, debemos hacer un par de cosas:

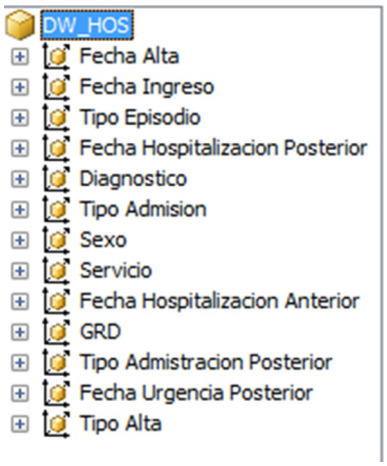
- Acabar de configurar correctamente el nombre de las dimensiones, puesto que en realidad tenemos trece dimensiones, cinco de estas apuntando a dimensiones ya existentes.

- Definir las jerarquías en las dimensiones.

Para cambiar el nombre de las dimensiones restantes, debemos fijarnos en el lateral izquierdo inferior, donde tenemos el listado de dimensiones existentes. Si hacemos clic en el botón derecho del ratón, tendremos acceso a un menú que nos permite cambiar el nombre de las dimensiones.



Cambiamos el nombre de las dimensiones hasta generar el resultado siguiente.

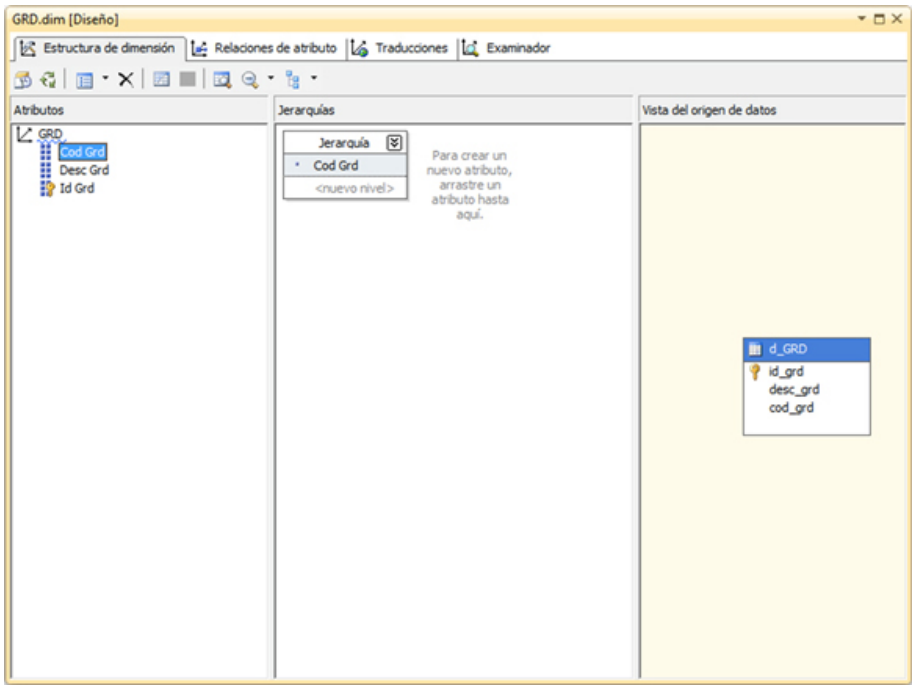


5) Personalizando las dimensiones y creando sus jerarquías

Queda, por último, definir las jerarquías para cada dimensión. Por defecto, el asistente de creación del cubo solo incluye la clave primaria como atributo accesible de la dimensión. Es necesario editar una por una cada dimensión, para indicar qué otros atributos queremos que sean accesibles.

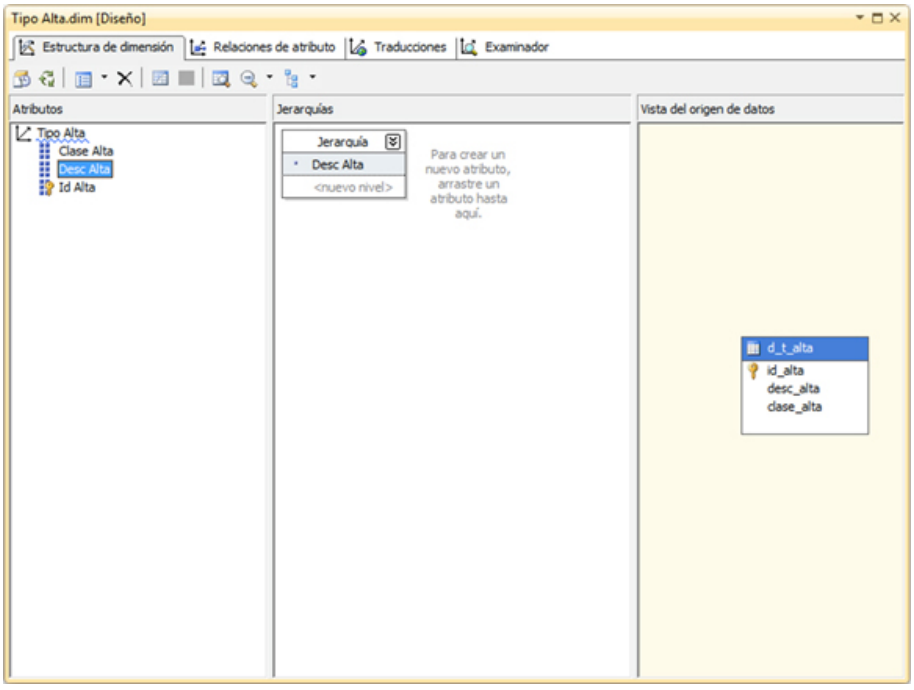
Las dimensiones que debemos editar las tenemos accesibles en el explorador de soluciones. Abrimos, por ejemplo, la dimensión GRD. Como ya se ha comentado, no tenemos ninguna jerarquía definida.

Para que la dimensión reconozca los atributos que faltan, debemos arrastrarlos desde la vista del origen de datos a la zona de atributos. Después, podemos usar los atributos para crear la jerarquía. En el caso de la dimensión GRD, nuestra jerarquía solo tiene un nivel. Por tanto, el resultado final quedará como se muestra en la figura siguiente.

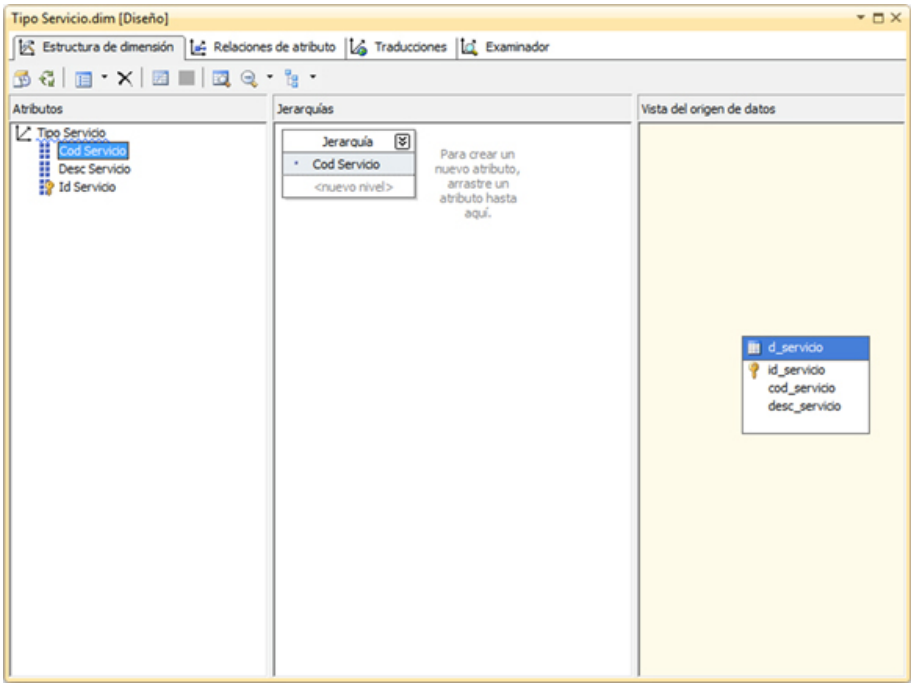


Este mismo proceso se debe hacer para cada una de las dimensiones. Las siguientes capturas muestran el diseño final de las jerarquías para cada dimensión:

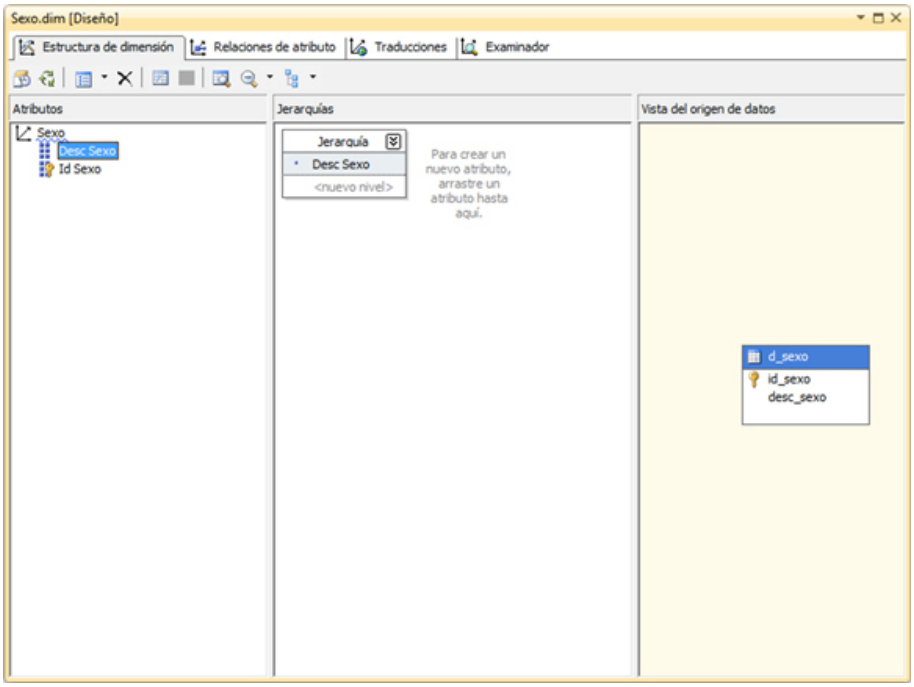
- Dimensión tipo alta



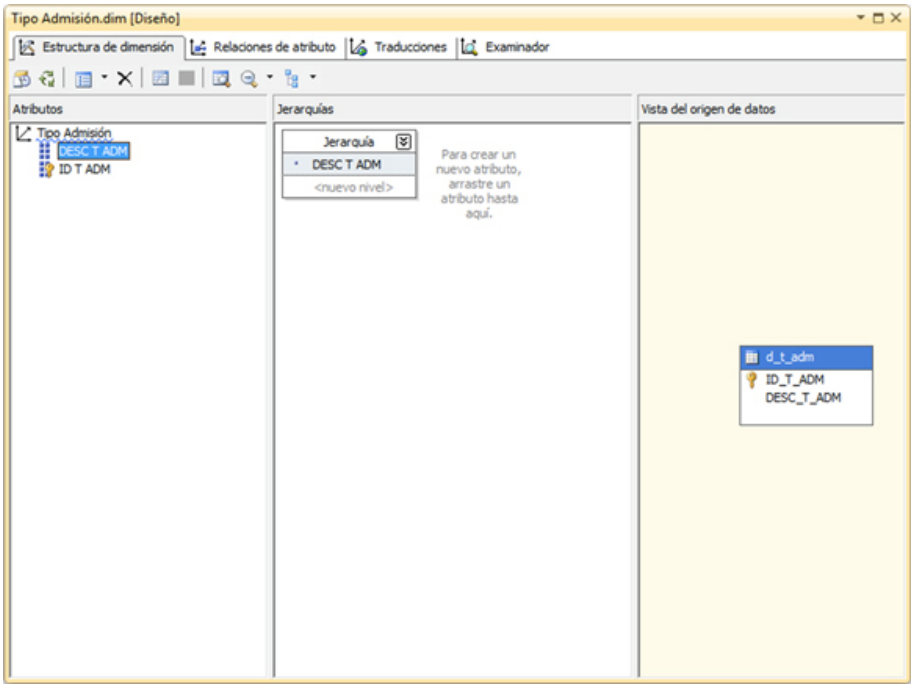
- Dimensión servicio



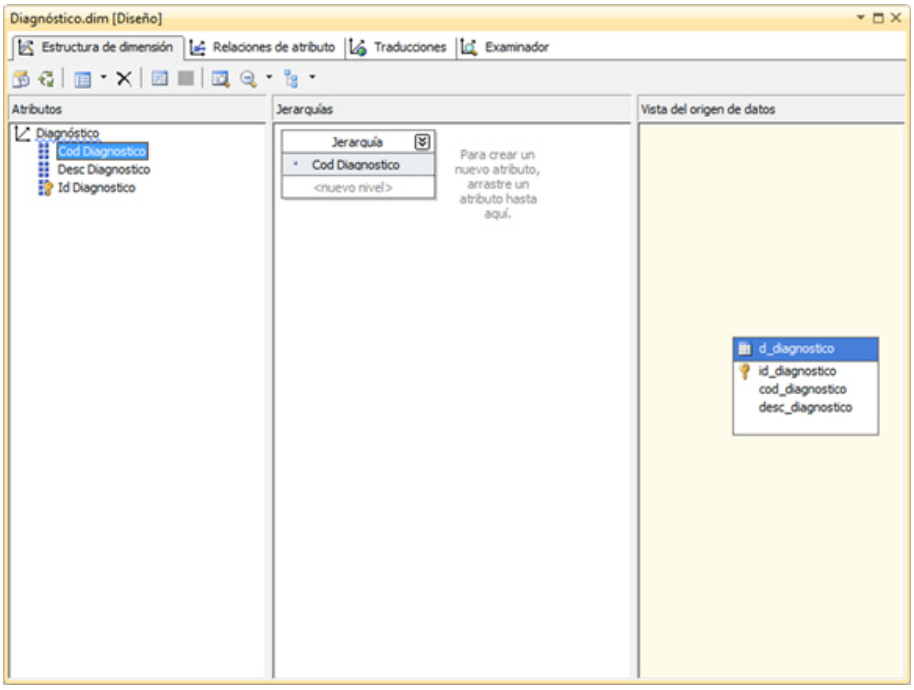
- Dimensión sexo



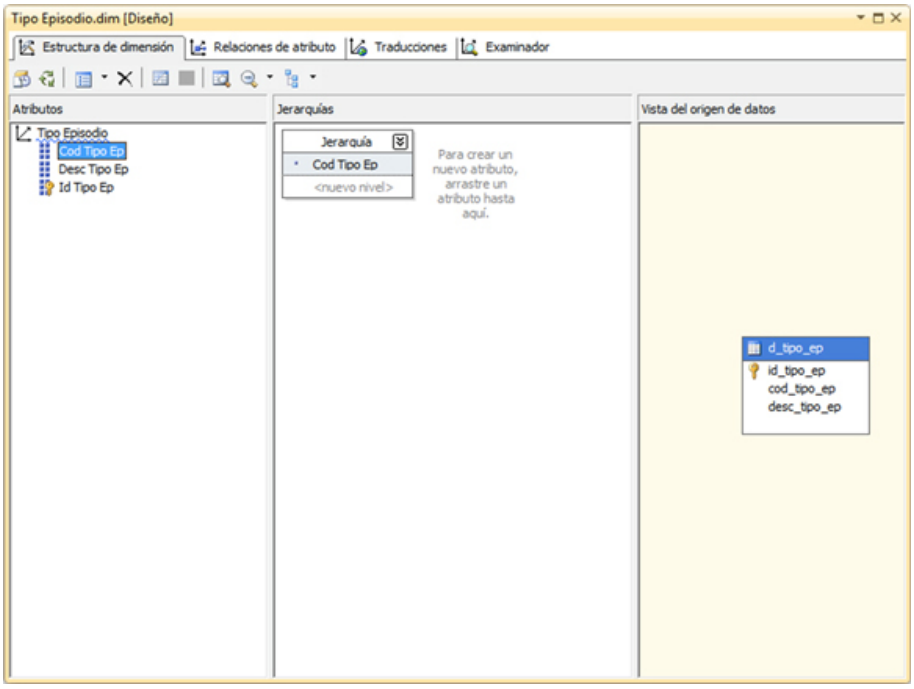
- Dimensión tipo admisión



- Dimensión diagnóstico



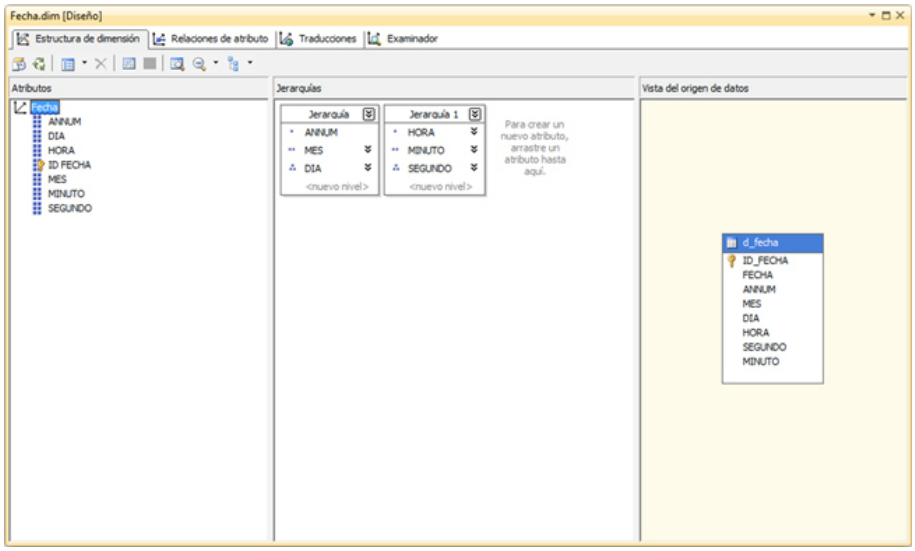
- Dimensión tipo episodio



- Dimensión fecha
  Tal y como se ha comentado, la dimensión temporal puede tener distintas jerarquías de navegación. Por ejemplo:
  - Año > Mes > Día
  - Año > Mes > Día > Hora > Minuto > Segundo
  - Año > Semana del año

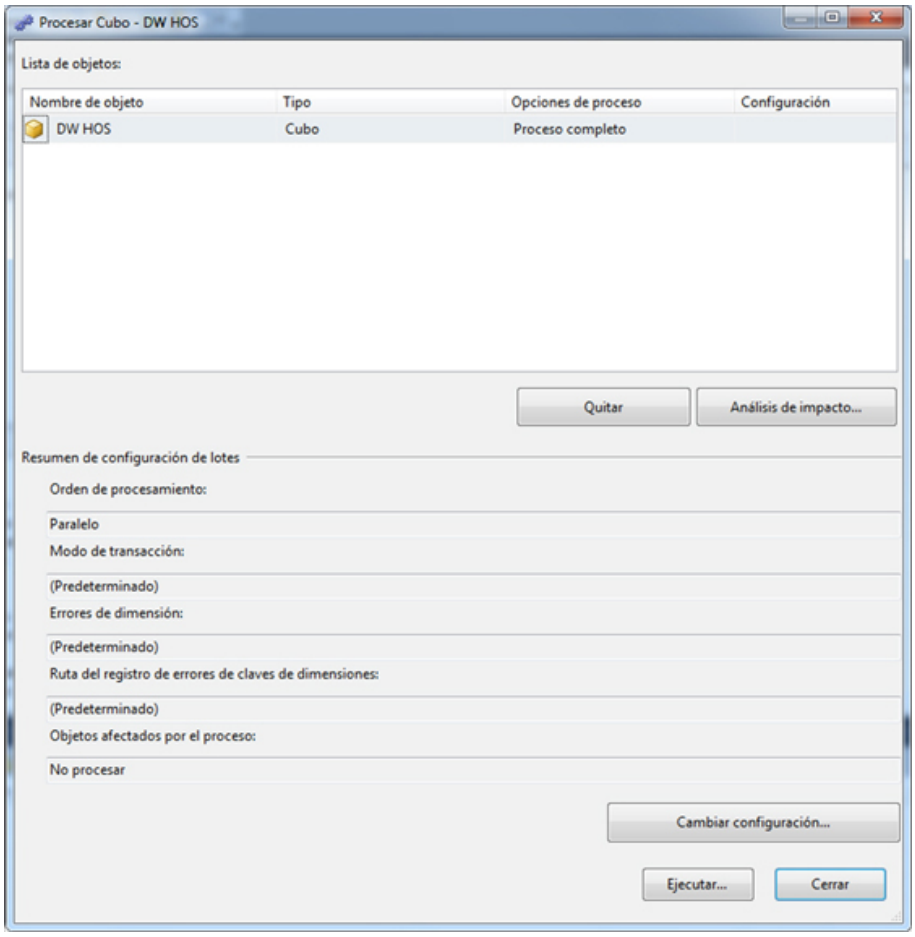
  En nuestro caso, hemos definido dos jerarquías:
  - Año > Mes > Día
  - Hora > Minuto > Segundo

En el caso de tener más atributos en la dimensión temporal como semana del año, día de la semana, trimestre o semestre, se podrían definir jerarquías más complejas.



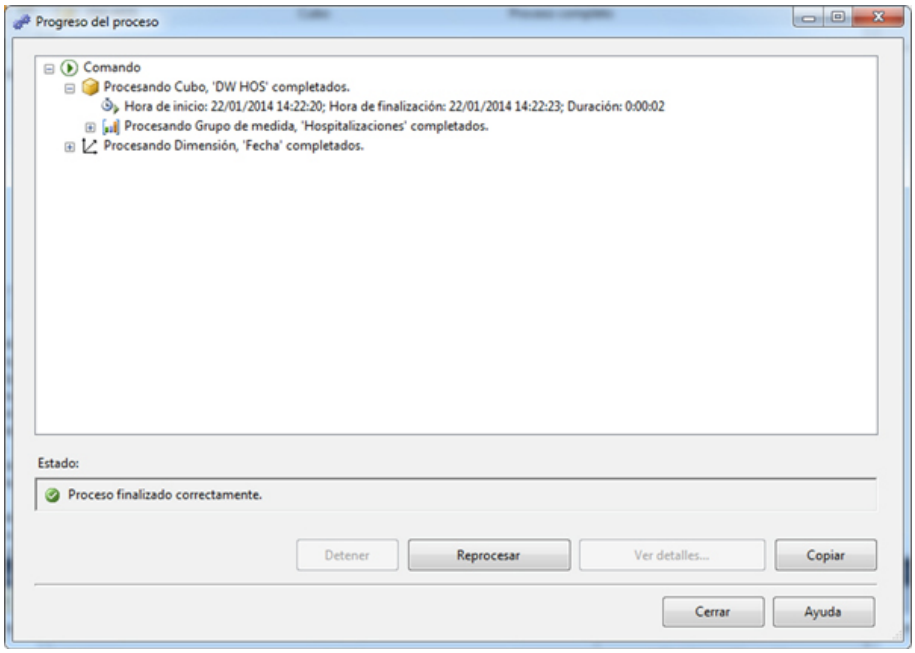
Este proceso de mejora del cubo aún puede refinarse más, por ejemplo renombrando los atributos y las jerarquías para que utilicen un lenguaje más cercano al negocio. Y de este modo, como resultado, facilitar su uso.

Por último, es necesario procesar el cubo. El procesado tiene dos partes. Por un lado, la validación del cubo y, por otro, la generación del cubo.



Ejecutamos este proceso de generación del cubo. Primero se validará[10] que la estructura creada es correcta, y posteriormente se generará su estructura. El resultado es:

(10)También se puede ejecutar la validación de manera independiente.



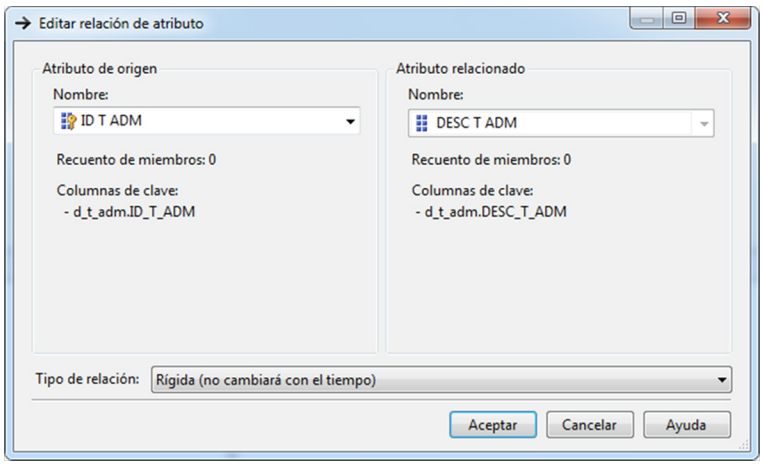
Es importante recordar que si hacemos cualquier cambio en el cubo, será necesario llevar a cabo un reprocesado del cubo.

6) Corrigiendo las alertas vinculadas al cubo

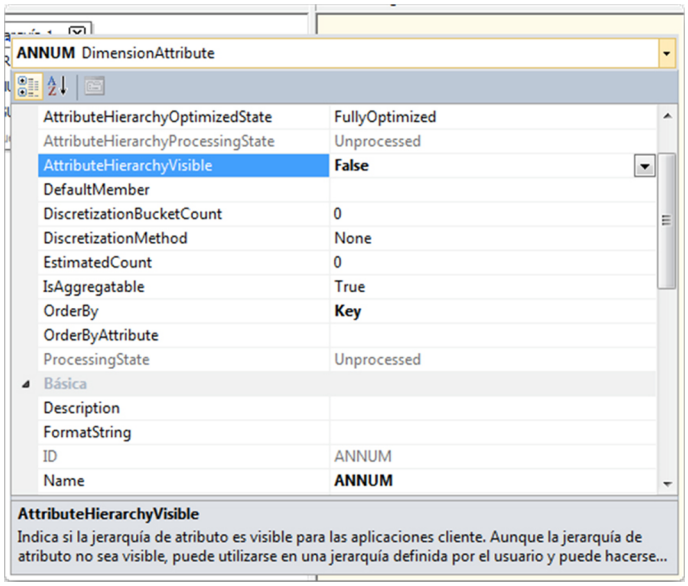
Tras la generación del cubo, es posible apreciar distintas alertas. Estas alertas son recomendaciones para el diseño del cubo. Son de dos tipos:

- Uso de relaciones rígidas en lugar de flexibles.
- Ocultación de atributos si se utilizan jerarquías.

Cada una de las relaciones creadas por defecto por el generador de dimensiones es flexible, lo que significa que las relaciones entre los miembros pueden cambiar con el tiempo. Como no es nuestro caso, lo que haremos es modificar todas las relaciones en cada dimensión y, de este modo, haremos desaparecer este tipo de alertas. Por ejemplo, en la dimensión tipo admisión.



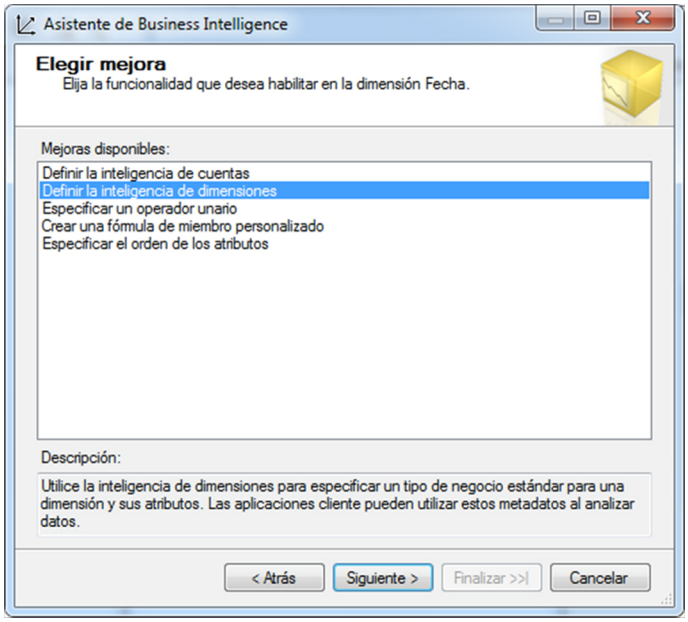
En el caso de que solo nos interese mostrar las jerarquías y no los atributos de manera independiente, debemos editar las propiedades de cada uno de los mismos cambiando su visibilidad a *false* mediante la propiedad *AttributeHierarchyVisible*.



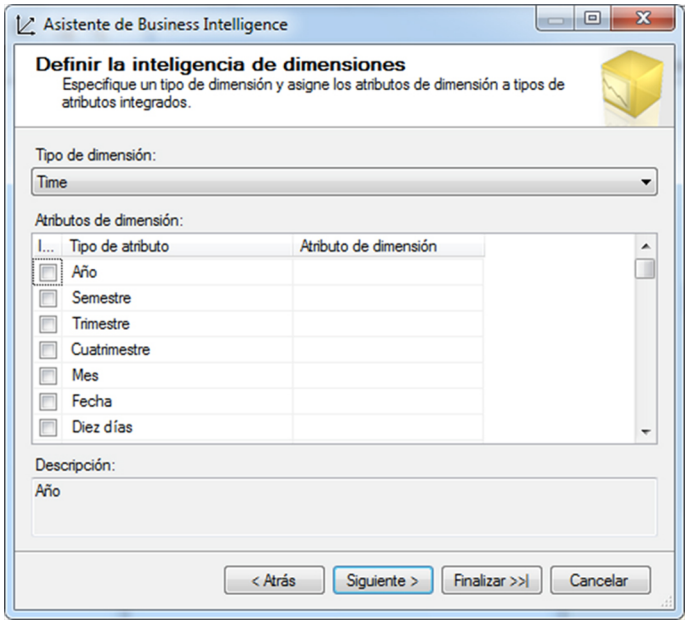
7) Mejorando la dimensión temporal

Una manera de mejorar la dimensión fecha consiste en indicar al sistema que esta dimensión es una dimensión temporal. Este tipo de acción se lleva a cabo mediante el asistente de *business intelligence* aplicado a esta dimensión.

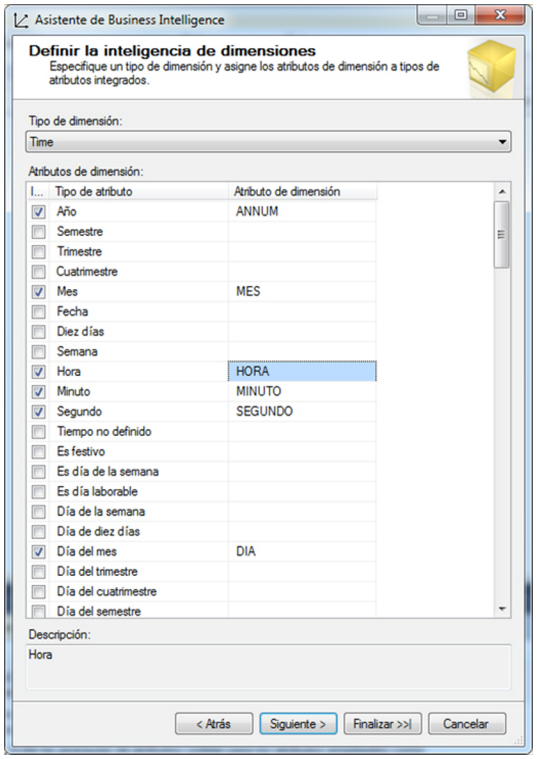
Indicamos que vamos a definir la inteligencia de la dimensión.



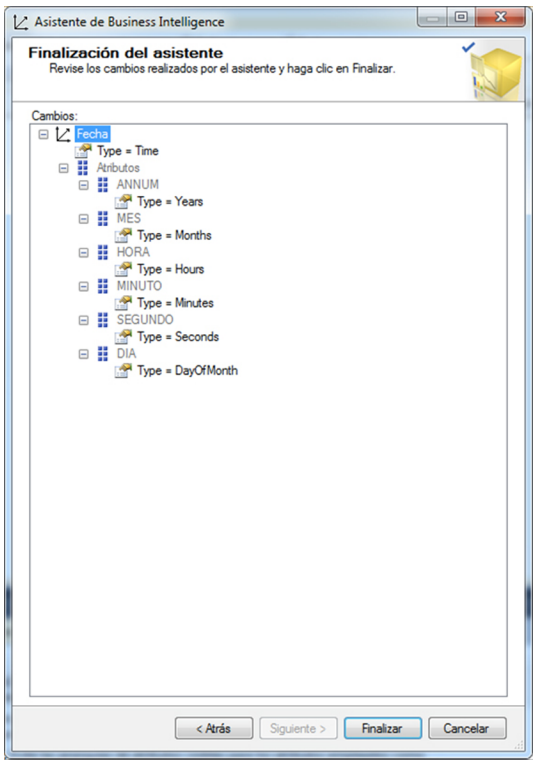
Como podemos apreciar, el editor presenta muchos tipos diferentes. Seleccionamos *Time*.



Teniendo en cuenta los atributos de nuestra dimensión, podemos indicar para cada uno de los mismos el tipo de atributo que es. Resulta importante ver que podíamos haber enriquecido más nuestra dimensión temporal (añadiendo, por ejemplo, el nombre de los meses y de los días, indicando qué días son festivos, etc.) mediante los procesos ETL. En nuestro caso particular:

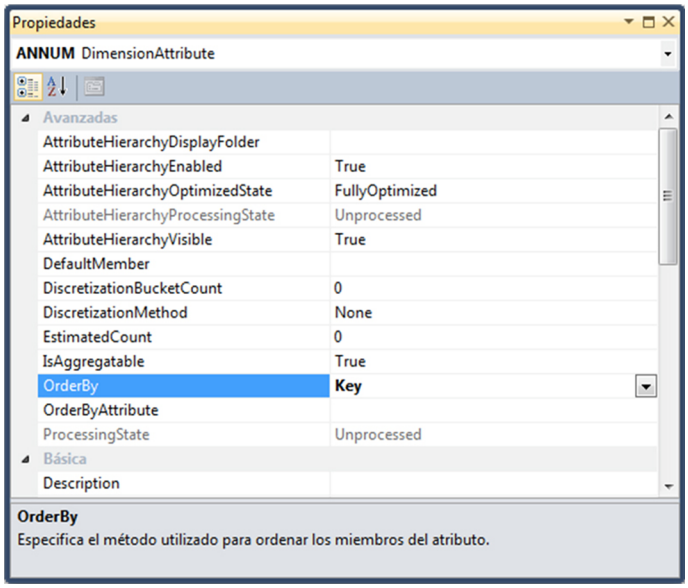


Como resultado, tenemos lo siguiente.



Será necesario reprocesar el cubo para que estos cambios estén disponibles.

Otro cambio importante que es necesario hacer cuando las dimensiones se han creado por defecto consiste en indicar el criterio de ordenación de los atributos. En el caso de la dimensión temporal, debemos editar las propiedades de cada elemento indicando que se ordenen por *key*, para que estén ordenados correctamente.



5.3. Tres vistas OLAP / respuestas

Tras la creación del cubo, ya estamos preparados para la creación de las consultas mínimas que han sido definidas por los usuarios de negocio. Como se indica en el enunciado de la actividad, el sistema debe ser capaz de responder a las preguntas siguientes.

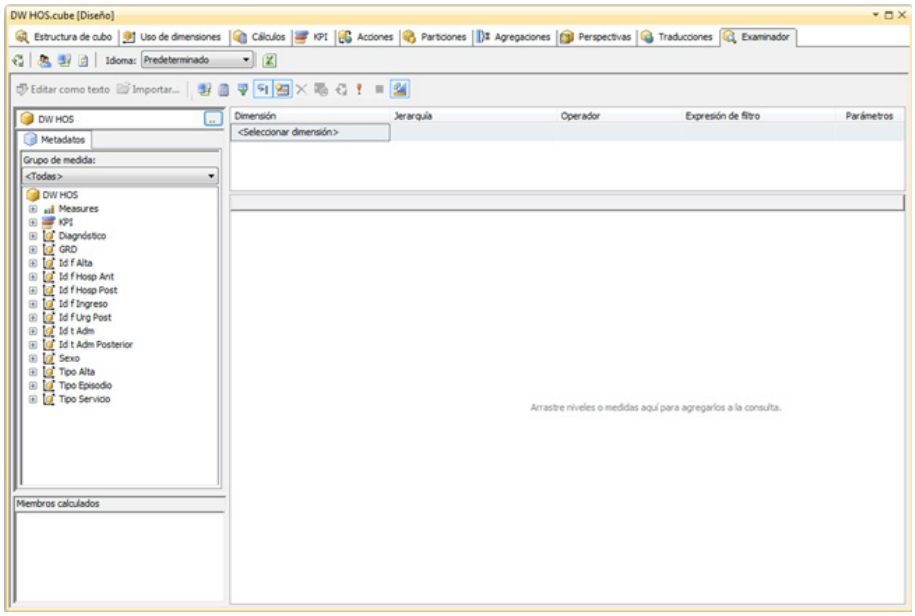
- Evolución de las hospitalizaciones
- Evolución de las hospitalizaciones por tipo de alta
- Evolución de las hospitalizaciones por meses y años
- Evolución de las hospitalizaciones por servicio del hospital

En este apartado final de la solución, veremos cómo el cubo OLAP que hemos diseñado es capaz de responder a cada una de las preguntas anteriores.

La consulta al cubo puede hacerse de diferentes maneras.

- Directamente, mediante la pestaña Examinador en la edición del cubo.
- Mediante un visor OLAP, como por ejemplo Microsoft Excel u otros.

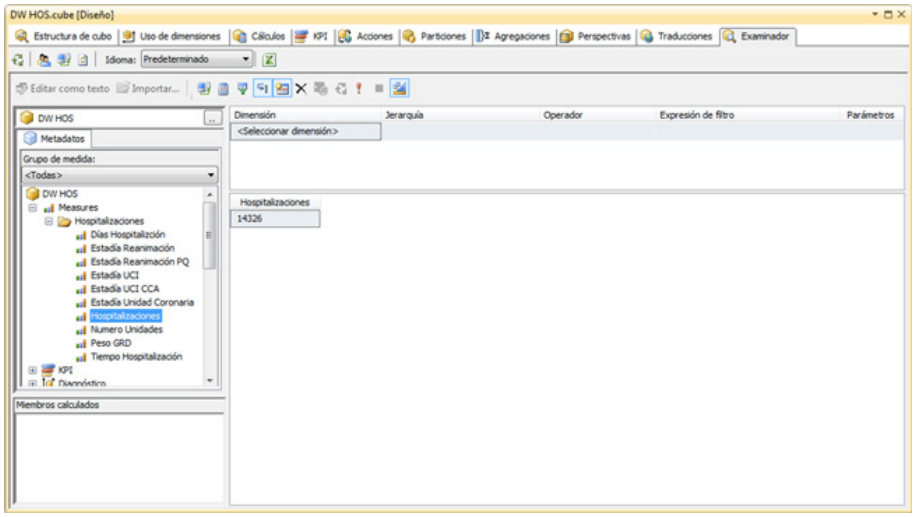
En nuestro caso particular, lo haremos directamente con el examinador.



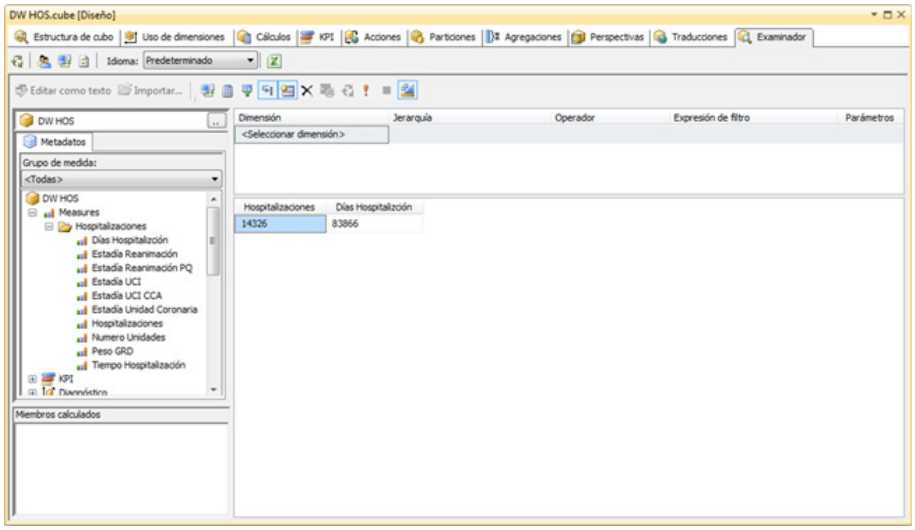
En caso de que las consultas sean sencillas, tan solo debemos arrastrar medidas o niveles para ser capaces de hacer preguntas al sistema. Vamos a revisar cada una de las preguntas.

1) Evolución de las hospitalizaciones

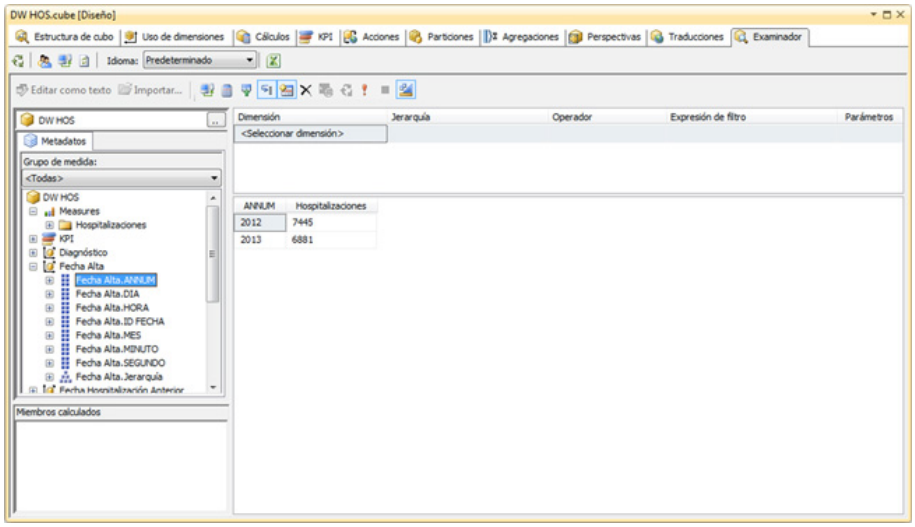
Si arrastramos la medida de hospitalizaciones rápidamente, sabremos que el número total es 14.326.



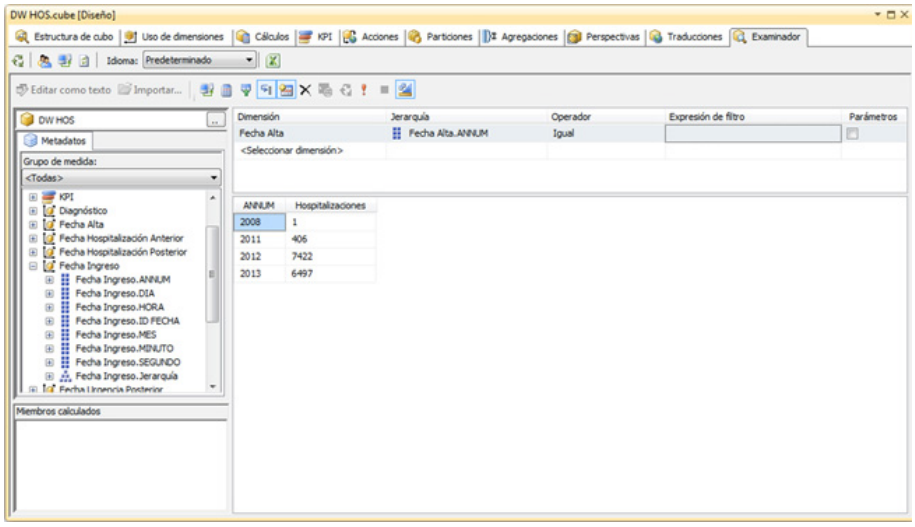
También podemos preguntarnos por los días de hospitalización totales.



Al añadir, por ejemplo, el nivel Año de la dimensión Fecha alta, descubrimos rápidamente que hubo más altas de hospitalizaciones en el 2012 que en el 2013.

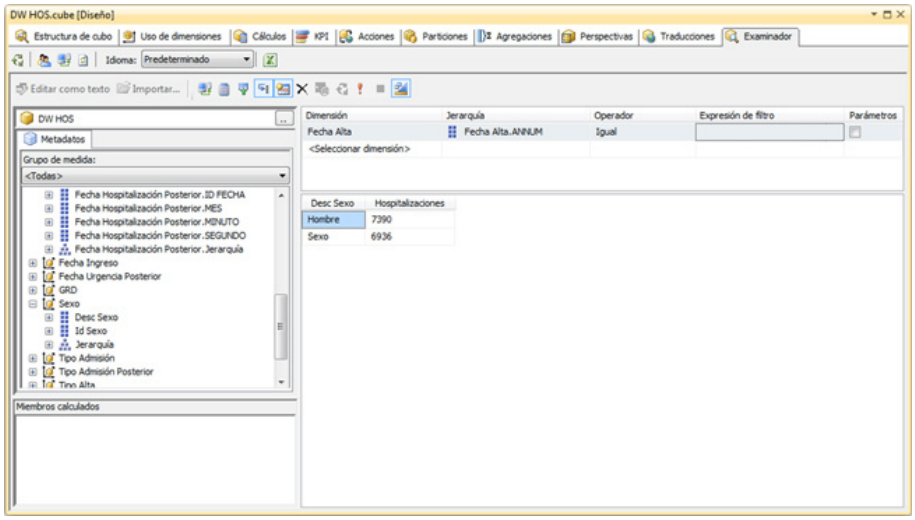


Y también podemos estudiar las hospitalizaciones por fecha de ingreso.

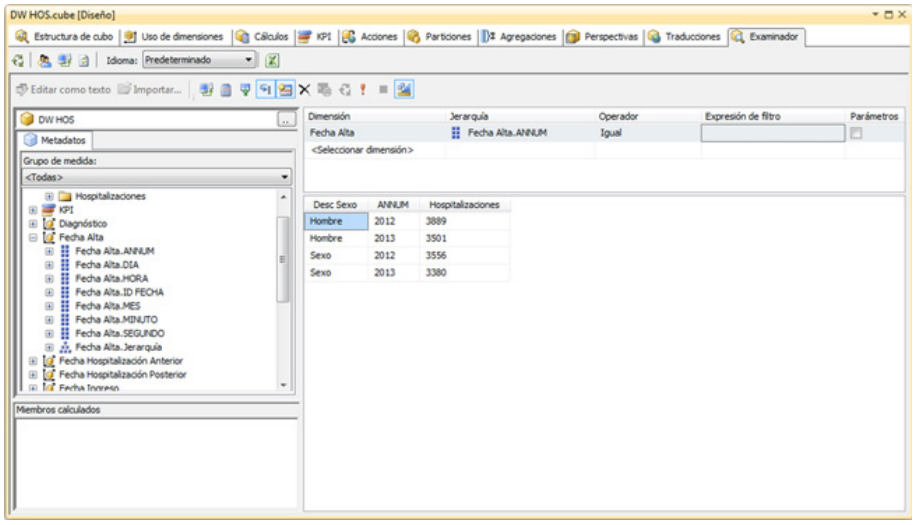


Como es posible apreciar para el conjunto de pacientes que estamos considerando, los ingresos de hospitalizaciones se concentran en el 2012 y el 2013.

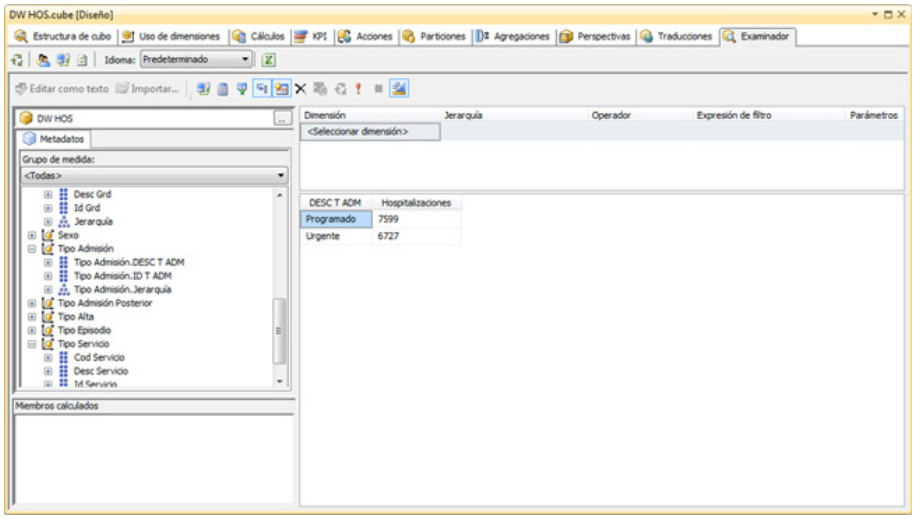
También podemos hacernos otras preguntas respecto a las hospitalizaciones como, por ejemplo, si ingresan más hombres que mujeres.



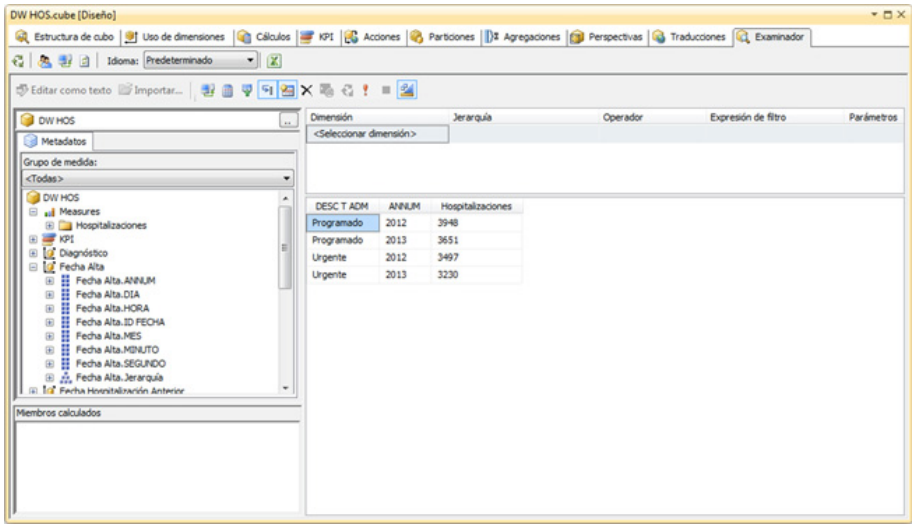
Parece que el número de hospitalizaciones de hombres es superior al de mujeres. Se puede consultar este hecho de manera histórica (por años):



Y también podemos preguntarnos si hubo más hospitalizaciones programadas o urgentes, a lo que el sistema nos responde:

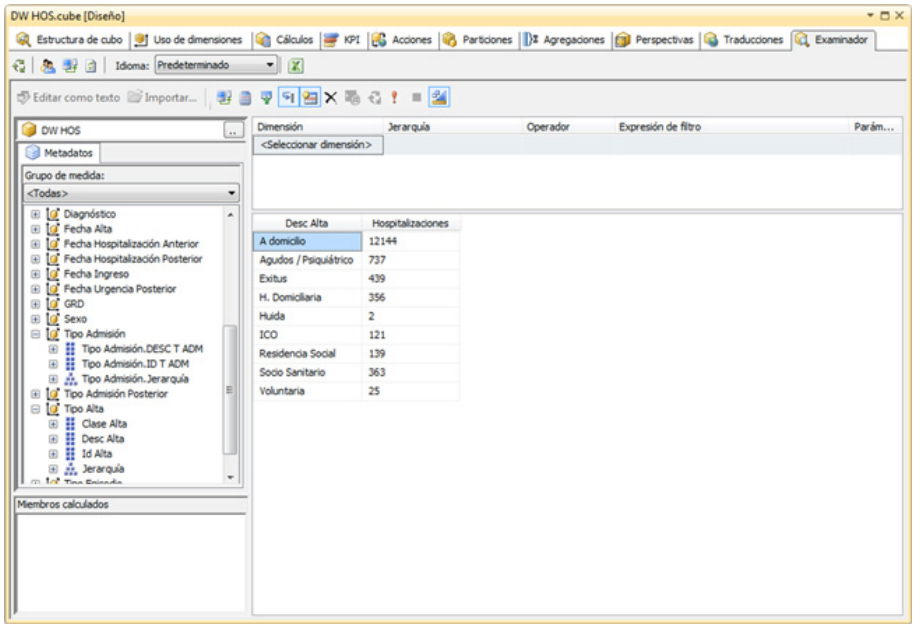


Y de nuevo, interesarnos por la evolución anual.



2) Evolución de las hospitalizaciones por tipo de alta

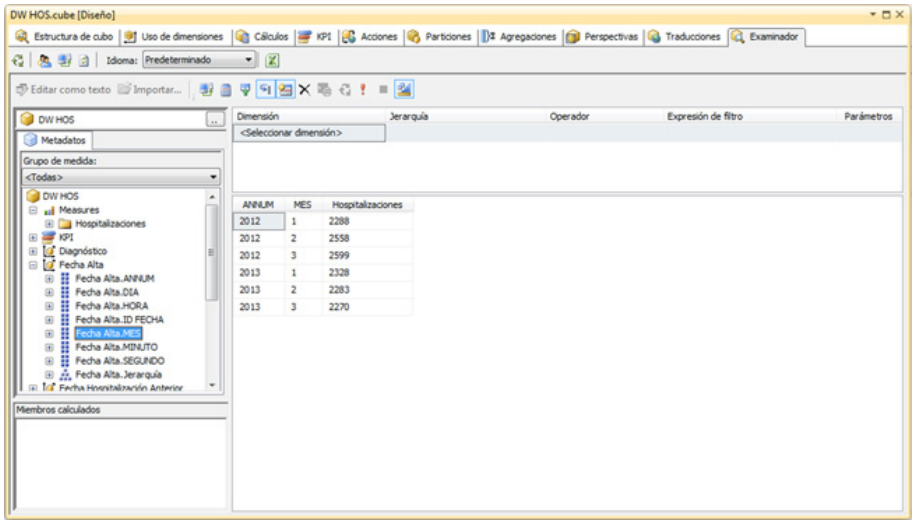
Si a la medida de hospitalizaciones le añadimos el nivel descriptivo del tipo de alta, veremos la distribución de la medida por cada uno de los valores:



Podemos comprobar, por lo tanto, que el tipo de alta más frecuente es a domicilio, seguida por agudos/psiquiátrico y *exitus*.

3) Evolución de las hospitalizaciones por meses y años

A la consulta anterior, en la que estudiábamos la evolución de las hospitalizaciones por fecha de alta, podemos añadir años y meses, como vemos a continuación:

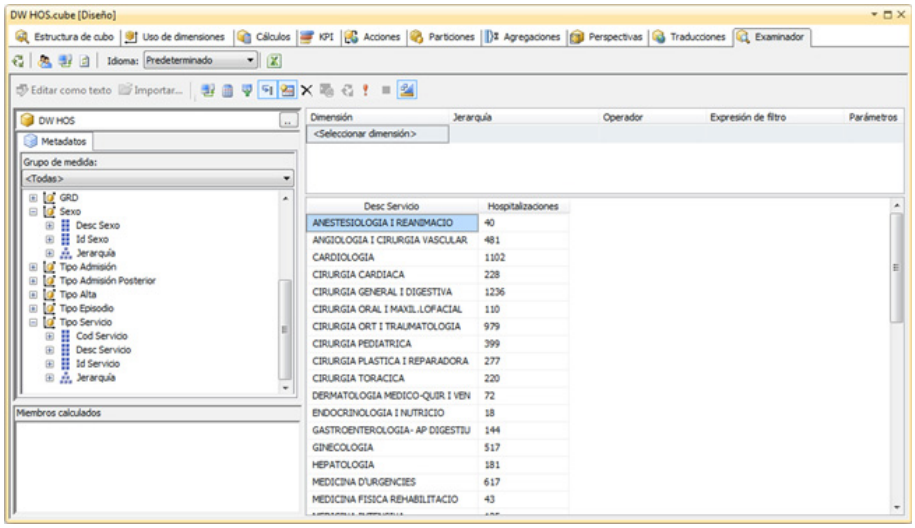


Por lo tanto, podemos apreciar lo siguiente.

- En enero del 2012 hubo menos hospitalizaciones que en enero del 2013.
- En febrero y marzo del 2013 hubo menos hospitalizaciones que en sus respectivos meses en el año anterior.

4) Evolución de las hospitalizaciones por servicio del hospital

Como en los casos anteriores, simplemente tenemos que arrastrar los niveles y las medidas que nos interesan.



De igual manera, añadiendo las fechas de ingreso podríamos estudiar cómo ha evolucionado el servicio.

Estas son algunas de las preguntas que puede hacerse un usuario. A medida que el usuario conozca más a fondo los datos presentes, así como sus capacidades, es probable que, de manera natural, pida mayor información. Esto implicará enriquecer el modelo, incluyendo nuevas métricas o más información para las dimensiones.