Metodología para el Diseño de Almacenes de Datos: Etapa de Modelado Conceptual

José María Cavero[1], Esperanza Marcos[1] y Mario Piattini[2]
[1] Escuela Superior de Ciencias Experimentales e Ingeniería, Universidad Rey Juan Carlos, Calle Tulipán s/n, 28933 Móstoles (Madrid), España, Fax: 34 91 6647490, {j.m.cavero,e.marcos}@escet.urjc.es
[2] Escuela Superior de Informática, Universidad de Castilla-La Mancha, Ronda de Calatrava 5, 13071 Ciudad Real, España, Fax: 34 926 295354, mpiattin@inf-cr.uclm.es

Abstract

El desarrollo de un Almacén de Datos (Data Warehouse) se ha convertido en un factor crítico de éxito para muchas compañías. De su calidad puede depender la supervivencia de la compañía en un mercado cada vez más competitivo. Por tanto, no es razonable manejar el proceso de construcción fuera del marco de trabajo de una metodología. Desgraciadamente, existen pocas metodologías completas para el diseño de almacenes de datos. En este trabajo se presenta MIDEA, una metodología basada en un modelo conceptual multidimensional. La metodología utiliza como marco de referencia la versión 3 de la Metodología Pública Española de Planificación y Desarrollo de Sistemas de Información (METRICA). Parte de la metodología está soportada por una herramienta CASE (IDEA-DWCASE), de la que se dispone de un primer prototipo. Ofreceremos una visión general tanto de la metodología como del modelo conceptual, y profundizaremos, a modo de ejemplo, en una de las actividades que la componen: el modelado conceptual.

1. Introducción

La necesidad de poder disponer de una forma rápida y sencilla de toda la información histórica presente en los sistemas operacionales y su uso para la toma de decisiones ha empujado a las empresas y a la comunidad científica a buscar nuevas formas de estructuración y acceso a estos datos de forma eficiente para, de esta forma, conseguir una ventaja con sus competidores. Existe un acuerdo en que los sistemas tradicionales de bases de datos no resultan adecuados para realizar consultas analíticas sobre ellos desde una perspectiva multidimensional, que es la forma en la que los analistas de negocio ven los datos de la organización. Los sistemas OLTP (*On-line transactional processing*) tradicionales están optimizados para proporcionar un elevado rendimiento en el procesamiento de un gran número de transacciones concurrentes, que habitualmente afectan a un reducido número de registros, mientras que los sistemas multidimensionales han de responder a consultas complejas (a veces impredecibles) que acceden a una enorme cantidad de registros [1].

Una posible solución consiste en la implantación de un sistema de almacén de datos, que proporciona un repositorio de información procedente fundamentalmente de sistemas operacionales (OLTP) que proporciona los datos para el procesamiento analítico y la toma de decisiones. Contiene datos refinados, históricos, resumidos y no volátiles, y son bases de datos fundamentalmente de sólo lectura, es decir, las actualizaciones se llevan a cabo esporádicamente, de forma controlada y masiva, y habitualmente fuera de los horarios de trabajo.

El interés de la industria en este tipo de tecnología se refleja en el hecho de que el crecimiento de las ventas y estimaciones realizadas entre 1998 y 2002 es superior al 20% anual [14].

Por lo tanto, los sistemas de almacén de datos proporcionan a los analistas un entorno integrado de información organizada de acuerdo a sus requisitos. Habitualmente se utilizan herramientas OLAP (*On-line Analytical Processing*) como herramientas frontales para el acceso a los datos. Aunque existen modelos híbridos, y distintas variaciones sobre los básicos, podemos hablar fundamentalmente de dos tipos de arquitecturas [15]: la arquitectura ROLAP (*Relational OLAP*), y la arquitectura MOLAP (*Multidimensional OLAP*).

En una arquitectura ROLAP, los datos se almacenan en tablas relacionales organizadas en un *esquema en estrella* [7] o alguna de sus variantes, ofreciendo de esta forma una interfaz multidimensional a las tablas relacionales. Un

esquema en estrella consiste en una tabla central de hechos de gran tamaño y varias tablas de dimensión a su alrededor cuyas claves primarias son claves ajenas en la tabla de hechos. Las medidas de interés para el proceso analítico se almacenan en las tablas de hecho, mientras que por cada dimensión (Tiempo, Geografía, etc.) existirá una tabla de dimensión, que contendrá todos los niveles de agregación (en el esquema en *copo de nieve*, o versión normalizada del esquema en estrella, cada nivel de agregación formará su propia tabla).

En una arquitectura MOLAP, sin embargo, los datos se almacenan directamente en estructuras multidimensionales, proporcionando por tanto directamente una visión multidimensional, sin ningún artificio similar al caso relacional. El rendimiento de este tipo de sistemas suele ser superior al caso relacional pero, probablemente debido a la poca madurez de este tipo de sistemas, todavía no son capaces de almacenar la gran cantidad de información que soportan los sistemas relacionales y por ello en ocasiones se utilizan como almacenes de datos departamentales (*data marts)* que pueden alimentarse de un almacén de datos corporativo relacional.

Los entornos OLTP y OLAP son profundamente diferentes, y las técnicas utilizadas para el diseño de bases de datos operacionales son inapropiadas para el diseño de almacenes de datos [7], [8]. El proceso de desarrollar un almacén de datos es, como cualquier tarea que implique algún tipo de integración de recursos pre-existentes (en este caso, datos procedentes fundamentalmente de sistemas heredados), sumamente complejo, y exigirá *"un gran esfuerzo, sujeto a errores, generalmente frustrante, y que lleva a que muchos proyectos se abandonen antes de su terminación"* [13].

A este respecto, en los últimos años ha habido bastantes propuestas restringidas a algunos de los aspectos particulares del diseño de los almacenes de datos, sin embargo, "*aunque se han desarrollado muchas soluciones para subproblemas interesantes, como el manejo de datos multidimensionales, mantenimiento de vistas para datos agregados, integración de datos, etc., la combinación de estas soluciones parciales y a menudo muy abstractas en una metodología completa de diseño y una estrategia de warehousing todavía se deja en manos de los desarrolladores*" [4].

En el siguiente apartado se resumen algunos trabajos relacionados. En el apartado 3 ofrecemos una visión general de la metodología. El apartado 4 muestra, a modo de ejemplo, un resumen de una de las actividades de la metodología. Por último, terminaremos con unas conclusiones.

2. Trabajos relacionados

A pesar de la evidente importancia que tiene disponer de un soporte metodológico para el desarrollo de un sistema OLAP de calidad, el proceso de diseño hasta ahora ha recibido muy poca atención por parte de la comunidad científica y de los proveedores de productos. Los modelos habitualmente utilizados para el diseño de bases de datos operacionales, como el modelo E/R, no deberían utilizarse sin más para el diseño de entornos analíticos. Atendiendo a motivos puramente técnicos, las bases de datos obtenidas como resultado del modelado con esta técnica son inapropiadas para sistemas de soporte a la decisión en los que es importante la eficiencia en las consultas y en la carga de los datos (incluyendo las cargas incrementales) [2]. Además, como se señala en [7], los modelos de datos E/R *"no son comprendidos por los usuarios y no puede navegarse de forma útil por ellos mediante el software de los SGBD"*. Por tanto, no sólo debería ser obligatorio que el paradigma multidimensional se utilizara para consultar la base de datos, sino que también debería utilizarse para su diseño y mantenimiento. Para utilizar el paradigma multidimensional durante todas las fases de desarrollo es necesario *"definir para este paradigma modelos de datos conceptuales, lógicos y físicos, y desarrollar una metodología válida que proporcione guías acerca de cómo crear y transformar estos modelos durante el proceso de desarrollo"* [3]. En [16] se propone la utilización del modelo multidimensional para la fase de modelado conceptual y el relacional para las fases de diseño lógico y físico, debido a su sólido fundamento matemático para el procesamiento de consultas, y reclaman la necesidad de metodologías y herramientas de diseño para almacenes de datos con un soporte apropiado para la jerarquías de agregación, correspondencias entre modelos multidimensionales y relacionales y modelos de coste para el particionamiento y la agregación que pueden utilizarse desde las primeras etapas del diseño.

En los últimos años han aparecido diferentes propuestas metodológicas para el desarrollo de almacenes de datos. Por ejemplo, en [8] los autores plantean una aproximación basada en dos puntos: por una parte, la Arquitectura en Bus del Almacén de Datos (*Data Warehouse Bus Architecture*), que mostrará cómo construir una sucesión de almacenes de datos departamentales que, finalmente, permitirán crear un almacén de datos corporativo y, por otra, la aproximación basada en el Ciclo de

Vida Dimensional del Negocio (*Business Dimensional Lifecycle -BDL- approach*), que tiene como objetivos la construcción, a partir de los requisitos del negocio, de almacenes de datos departamentales basados en modelos dimensionales en estrella. Es una metodología muy detallada y, según los propios autores, ampliamente probada. Sin embargo, está excesivamente centrada en el modelo relacional ya desde las fases iniciales del modelado dimensional.

En [1], los autores presentan tanto un modelo lógico para el diseño de bases de datos multidimensionales (llamado MD) como una metodología de diseño para obtener un esquema MD a partir de bases de datos operacionales. Para ello utilizan como punto de partida un esquema E/R que describe una vista integrada de las bases de datos operacionales, que contendrá toda la información disponible para nuestro almacén de datos, aunque en un formato no adecuado a este tipo de sistema. La metodología consta, por una parte, de una serie de pasos para la construcción del esquema en el modelo MD, y por otra de una transformación tanto al modelo relacional como a matrices multidimensionales. La metodología es aún incompleta y parte de una situación ideal, suponiendo que toda la información estará incluida en el esquema E/R. Sin embargo, los esquemas operacionales deberían ser simplemente un apoyo, dando una mayor importancia a los requisitos de los usuarios analíticos.

En [6] se esboza un marco metodológico para el diseño de almacenes de datos basado en el modelo conceptual de los mismos autores, llamado Dimensional Fact Model (DFM). La metodología aún se encuentra incompleta, y de momento se centra únicamente en la implementación relacional.

Existen muchas otras propuestas parciales, centradas en aspectos muy particulares como transformación entre modelos, materialización de vistas, índices, etc. Por ejemplo, en [12] se propone utilizar técnicas de data mining en las fases de diseño del almacén de datos (algoritmos de data mining para descubrir información implícita en los datos, para la resolución de conflictos de la integración de esquemas para la compleción de valores perdidos y la corrección de ruido en los datos y datos incorrectos, etc.).

Como resumen, podemos decir que aunque existe un acuerdo en cuanto a la necesidad de metodologías y herramientas para el desarrollo de almacenes de datos de calidad, todavía no existe ninguna unánimemente aceptada. En el presente artículo presentamos MIDEA, una metodología de desarrollo de almacenes de datos. La metodología utiliza en su fase de análisis un modelo conceptual multidimensional de datos, denominado IDEA, que permite modelar esquemas conceptuales multidimensionales.

3. Propuesta

Según [17], la calidad total de los sistemas de información es un concepto multidimensional, que engloba a las siguientes dimensiones (figura 1):

- Calidad de las infraestructuras, que engloba al hardware y el software que lo soporta (por ejemplo, redes, software de sistema, etc.)
- Calidad del software, es decir, la calidad de las aplicaciones construidas, mantenidas o soportadas por el departamento de Sistemas de Información.
- Calidad de los datos de entrada a los distintos sistemas de información.
- Calidad de la información, es decir, la calidad de las salidas resultantes de los sistemas de información. En ocasiones, la salida de un sistema se convierte en entrada de otro, por lo que la calidad de la información está relacionada con la calidad de los datos.
- Calidad administrativa, es decir, la calidad de la gestión en la función del departamento de Sistemas de Información.
- Calidad de los servicios, que incluye la calidad de los procesos de soporte al cliente, tales como los relativos a los "help desk".

La metodología de desarrollo que presentamos en el presente artículo tendría como objetivo fundamental conseguir la calidad de la información analítica suministrada a los usuarios que toman las decisiones en la empresa aunque, evidentemente, también influirá en (y se verá influida por) el resto de las dimensiones que engloban la calidad, tal como se muestra en la figura 1.

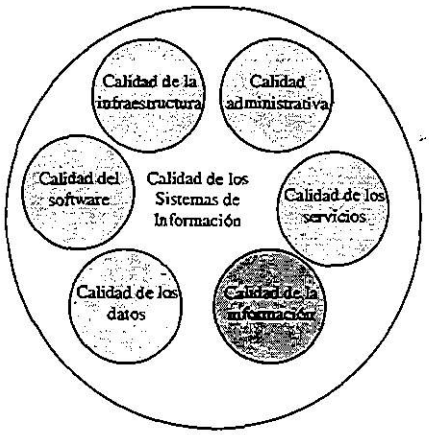

Figura 1. Dimensiones de la calidad

La metodología desarrollada se engloba dentro del marco del proyecto EINSTEIN. EINSTEIN es un proyecto de Investigación y desarrollo que aplica la experiencia y el conocimiento obtenido en el desarrollo de sistemas de bases de datos relacionales en la última década (SQL, modelado E/R, herramientas CASE, metodologías, ...) al diseño de bases de datos multidimensionales (BDMDs).

El proyecto propone una metodología de desarrollo de BDMD (Bases de Datos MultiDimensionales) análoga a las tradicionales que se han utilizado en el desarrollo de sistemas de bases de datos relacionales. La metodología utiliza como modelo conceptual en su fase de análisis un modelo conceptual multidimensional denominado IDEA, desarrollado asimismo en el marco del proyecto EINSTEIN [11]. Además, parte de esta metodología está soportada por una herramienta CASE (IDEA-DWCASE) que incorpora una interfaz gráfica [10]. Esta herramienta permite la transformación de un esquema conceptual IDEA en un esquema lógico basado en el modelo soportado por algunos productos multidimensionales o relacionales. La figura 2 muestra una ventana del prototipo de la herramienta, cuya notación gráfica se basa en [5].

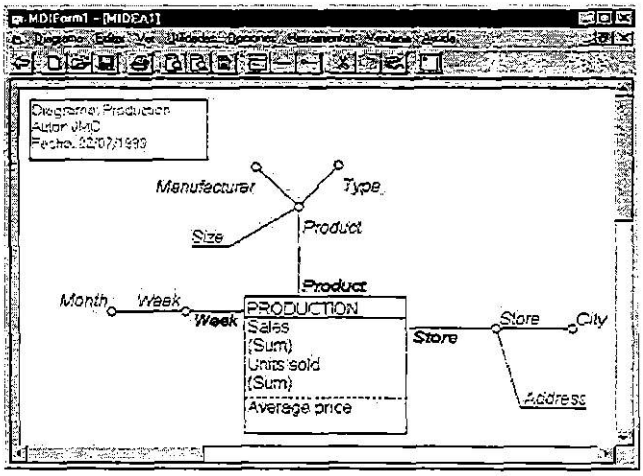

Figura 2. Prototipo IDEA-DWCASE

IDEA es el modelo de datos multidimensional utilizado como modelo conceptual para captar la semántica durante el desarrollo [11]. La estructura principal de IDEA es el Esquema de Hecho, y es similar a los conceptos de esquema de relación en el modelo de datos relacional y al de tipo de entidad en el modelo ER. Un Esquema de Hecho es la descripción de un espacio n-dimensional que contiene información relevante para su procesamiento analítico. Todo Esquema de Hecho consta de un conjunto de dimensiones, una estructura de celda, y puede, o no, tener definido un predicado, de tal forma que los datos contenidos en la extensión del EH sean únicamente aquellos que cumplan el predicado. Cada dimensión está asociada a un atributo de dimensión, el cual, en caso de que aquélla esté asociada a una subjerarquía de atributos, será su raíz. Por otra parte, la estructura de celda consta de una estructura de subcelda (que contiene un atributo de síntesis y un conjunto de funciones de síntesis definidas sobre éste) y puede tener un conjunto de métodos, que son procedimientos aplicados sobre una o más estructuras de subceldas.

A continuación ofreceremos una visión general de la metodología y profundizaremos, a modo de ejemplo, en una de sus actividades.

IDEA se utiliza para comprender y representar los requisitos de los usuarios analíticos de una forma similar a como el modelo de datos E/R se utiliza para interactuar con los usuarios de los microdatos. Los esquemas de datos elementales de los sistemas OLTP existentes y los requisitos obtenidos de los usuarios de datos analíticos son las entradas principales en la construcción de los esquemas conceptuales multidimensionales en IDEA.

El siguiente paso consiste en transformar, utilizando un conjunto de reglas metodológicas, cada esquema conceptual definido previamente en un esquema lógico en el modelo de cada producto concreto, el cual puede ser un sistema de gestión de bases de datos multidimensional puro o un sistema relacional con características multidimensionales (star joins, índices bitmap, ...). Es necesario destacar que el procedimiento habitual en los proyectos actuales es transformar directamente los esquemas relacionales en esquemas multidimensionales soportados directamente por herramientas OLAP.

La aproximación del proyecto EINSTEIN permite la ingeniería inversa de esquemas multidimensionales específicos existentes en esquemas conceptuales IDEA. Estos podrán comprobarse con los requisitos de los usuarios OLAP para verificar que el almacén de datos actual los satisface.

De igual modo, al contrario de la mayoría de las aproximaciones actuales, es posible crear y/o verificar esquemas E/R conceptuales elementales utilizando un conjunto de reglas contenidas en la metodología para satisfacer los requisitos de los usuarios analíticos. Creemos que esta aproximación no ha sido tratada en suficiente profundidad en los trabajos previos, en los que normalmente sólo podemos ver una dirección en el modelado dimensional: el que va desde las bases de datos operacionales hacia las analíticas, pero no el opuesto, es decir, desde las necesidades analíticas hacia un diseño operacional.

La metodología utiliza como marco de referencia la propuesta para la versión 3 de la Metodología Pública Española MÉTRICA (MV3) [9]. Los procesos pertenecientes a MV3 que se contemplan son aquéllos en los que la especificidad del desarrollo de un Almacén de Datos tiene una mayor influencia, en concreto, los procesos de Análisis del Sistema de Información (ASI), Diseño del Sistema de Información (DSI) y Construcción del Sistema de Información (CSI). Los nuevos procesos, modificados a partir de los propuestos en MV3, han recibido el nombre de ASI-MD (MultiDimensional), DSI-MD y CSI-MD. El considerar únicamente estos tres procesos no significa que el resto de los contemplados en MV3 no deban tenerse en cuenta en el desarrollo de un almacén de datos. Es evidente que existirá una implantación del sistema, que puede llevarse a cabo un estudio de viabilidad previo, etc., sin embargo, se ha considerado que no serían significativas las diferencias con respecto al desarrollo de cualquier otro sistema de información.

La figura 3 muestra una visión global de la metodología, mostrando el alcance de los tres procesos que la componen, ASI-MD, DSI-MD y CSI-MD.

Cada uno de estos procesos se divide en actividades y, a su vez, éstas se descomponen en tareas. El orden asignado a las actividades no debe interpretarse necesariamente como una secuencia en su realización, ya que éstas pueden realizarse en orden diferente a su codificación o bien en paralelo, intercalando tareas de actividades diferentes. Sin embargo, no se dará por concluido un proceso hasta no haber terminado todas sus actividades. En los gráficos que acompañan a cada proceso, se destacan las actividades que tengan una implicación destacada en el desarrollo de un almacén de datos.

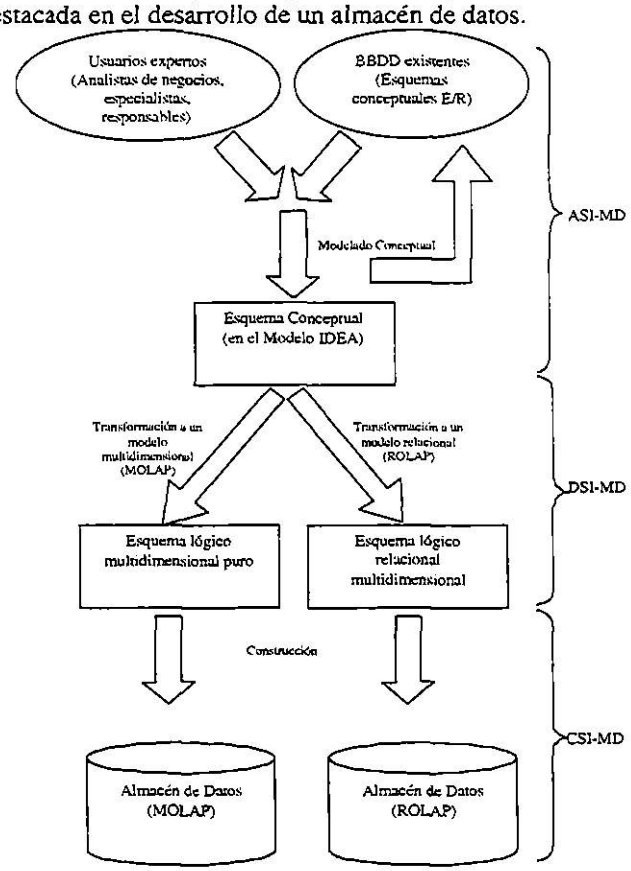

Figura 3. Visión general de la metodología

4. Ejemplo de Actividad: Modelado Conceptual del Almacén de Datos

En este apartado resumiremos las tareas que se han de llevar a cabo en la actividad de modelado conceptual del almacén de datos. Por falta de espacio, no se incluyen los productos de entrada y salida de cada tarea, los participantes en la misma y las técnicas utilizadas.

El objetivo de esta actividad consiste en obtener, aplicando el modelo IDEA, el esquema conceptual multidimensional de datos del almacén de datos, a partir del catálogo de requisitos y de los esquemas entidad/interrelación existentes. Consta de las siguientes siete actividades:

4.1. Obtención Preliminar de Estructuras de Subcelda

El primer paso en la obtención del esquema conceptual multidimensional es la obtención de un esquema preliminar, que se realizará en las dos primeras tareas.

Esta tarea consiste en la obtención de estructuras de subcelda preliminares. Estas estructuras de subcelda modelarán los "eventos que ocurren dinámicamente en el mundo de la empresa" [5], tales como las ventas en una empresa, los movimientos de una cuenta de ahorro, etc. No es necesario en este punto detallar cada estructura de subcelda hasta el nivel del atributo concreto y las funciones de síntesis que la forman. En este momento estamos interesados únicamente en estructuras generales, preliminares. La información necesaria para el modelado de las estructuras de subcelda preliminares puede proceder de distintas fuentes: por una parte, y más importante, la opinión de los usuarios expertos. Son ellos los que mejor conocen su problemática y por tanto los que saben qué datos son los que necesitan en su trabajo. Habitualmente se corresponden con medidas numéricas de la empresa, preferentemente valores "numéricos, continuos, y aditivos" [7]. Por otra parte, si contamos con el esquema entidad/interrelación de la base de datos de la empresa, estas estructuras de subcelda preliminares suelen corresponderse con determinados atributos de entidades o de interrelaciones N:M. También podremos contar con el esquema conceptual multidimensional previo elaborado en la actividad ASI-MD 1.

4.2. Obtención Preliminar de las Dimensiones

Una vez detectadas las estructuras de subcelda preliminares, que representan las variables de interés en la empresa, el siguiente paso consiste en detectar las dimensiones que formarán parte de cada uno de ellos, es decir, de qué forma se podrán agregar los valores que hemos detectado. En este momento los usuarios deben pensar en las dimensiones de forma abstracta, sin intentar necesariamente identificar los atributos que formarán su jerarquía de forma individual. Por ejemplo, los usuarios pensarán en dimensiones tales como el tiempo, el espacio, etc., y no en términos de atributos como días-meses-años (en el caso del tiempo) o delegación-provincia-país (en el caso del espacio). Esta forma de trabajar descendente puede complementarse con el estudio de los esquemas conceptuales de las bases de datos operacionales, observando los atributos que componen las entidades e interrelaciones que se encuentran enlazados (bien directamente, bien a través de otras) a aquéllas que se han identificado como hechos en el esquema entidad/interrelación. Estos atributos nos darán pistas acerca de posibles dimensiones ocultas no detectadas por los usuarios. Si contamos con un esquema general multidimensional como salida de la primera actividad del proceso, también podremos utilizarlo como fuente de información.

4.3. Obtención Preliminar de las Jerarquías

Llegados a este punto, contaremos con un esquema preliminar que constará de un conjunto de estructuras de subcelda preliminares definidas sobre determinadas dimensiones. Es el momento de intentar identificar de forma un poco más precisa las dimensiones y las correspondientes jerarquías. Se identificará cada dimensión, describiendo, en caso de que exista, la subjerarquía que la forma y las agregaciones de que consta ésta. No es necesario en este momento entrar en detalle de los dominios de dimensión de que consta cada agregación, ni en el detalle de las funciones de agregación. En este momento podrán identificarse nuevas dimensiones. Un ejemplo típico de éstas es el tiempo, que en ocasiones no está presente en las bases de datos elementales pero que suele ser una dimensión fundamental en todos los almacenes de datos.

4.4. Obtención Detallada de las Jerarquías

El siguiente paso consiste en depurar las jerarquías obtenidas en el paso anterior. Esta depuración consistirá en la enumeración detallada de los atributos de dimensión de que consta la subjerarquía asociada a la dimensión. Podrán detectarse atributos nuevos o eliminar otros inútiles, distinguir o añadir aquéllos que serán únicamente propiedades de los atributos propios de la jerarquía (atributos de descripción). Por ejemplo, en una dimensión que contemple las distintas delegaciones

de una empresa en un país, podemos contemplar el teléfono o la dirección de cada delegación. Estos atributos no son más que propiedades (atributos de descripción) del propio atributo de dimensión. Para todos los atributos se definirá su dominio, o se asignará a un dominio ya definido. Asimismo, se elaborará la jerarquía de agregaciones de dominios, especificando en mayor detalle las funciones de agregación entre los mismos.

4.5. Obtención Detallada de las Estructuras de Subcelda

Una vez obtenidas por completo las dimensiones y definidas sus jerarquías, pasaremos a estudiar en más detalle las estructuras de subcelda. Especificaremos para cada una de ellas el atributo de que consta y las funciones de síntesis que se aplicarán sobre el mismo. Se estudiará cada una de estas funciones con respecto a cada una de las dimensiones, comprobando que cada función de síntesis se puede aplicar a lo largo de cada una de las dimensiones (o atributo de dimensión), señalando aquéllas que no sean aplicables (aunque algunas son evidentes, por ejemplo no tiene sentido sumar la temperatura a lo largo de los días del año, en general deberán indicarlas explícitamente los usuarios). Habitualmente estas funciones de síntesis consistirán en la función suma, aunque en ocasiones el usuario puede desear otras distintas (por ejemplo, la media, el valor máximo, el mínimo, etc.).

4.6. Obtención de los Esquemas de Hecho

En este momento tenemos identificadas estructuras de subcelda, cada una con sus dimensiones asociadas. El siguiente paso consiste en la agrupación de las estructuras de subceldas en estructuras de celda, para formar esquemas de hecho. Como cada estructura de celda pertenece a un esquema de hecho, y a su vez cada esquema de hecho consta de unas dimensiones determinadas, deberemos detectar aquellas estructuras de subcelda que sean susceptibles de unirse. Esta unión puede consistir en:

Unión de dos estructuras de subcelda en una única, debido a que representan al mismo hecho (es decir, estaban duplicadas). Sus atributos de síntesis y sus dimensiones coincidirán. Las funciones de síntesis resultado de la unión serán la unión de las funciones de síntesis de ambas estructuras de subcelda, eliminando aquellas que estuvieran duplicadas. Se revisará la aplicabilidad de las mismas a las dimensiones o atributos de dimensión.

Agregación de dos estructuras de subcelda en una estructura de celda, o una estructura de subcelda a una estructura de celda ya identificada previamente. En este caso sus dimensiones deben tener cierta coincidencia. Decimos cierta coincidencia porque en ocasiones puede interesar unir estructuras en las que no exista una coincidencia total de dimensiones. En ese caso, habrá que estudiar cómo se comportan las agregaciones en el caso de las nuevas dimensiones (o atributos de éstas) incorporados. Es decir, que sean o no agregables en las mismas. Naturalmente, siempre tendremos la posibilidad de incorporar todas las estructuras de subcelda en una misma estructura de celda (esquema de hecho), pero en este caso puede ocurrir que existan numerosos atributos de síntesis que no sean agregables en muchas dimensiones. El esquema, además, puede resultar poco legible.

4.7. Verificación y Validación del Esquema Multidimensional

En esta tarea se verifica y valida el esquema conceptual multidimensional, con el objetivo de asegurar que el mismo sea completo, ajustado al catálogo de requisitos, y que siga unos criterios de calidad predeterminados.

Se realizarán otro tipo de comprobaciones, como por ejemplo, que los datos del almacén de datos que obtendremos de las BBDD elementales están disponibles en las mismas (si no están disponibles en dichas BBDD habrá que plantear, por ejemplo, la necesidad de modificarlas).

La figura 4 muestra un ejemplo de esquema resultante producto de la actividad anterior, cuya representación gráfica aparece en la figura 2.

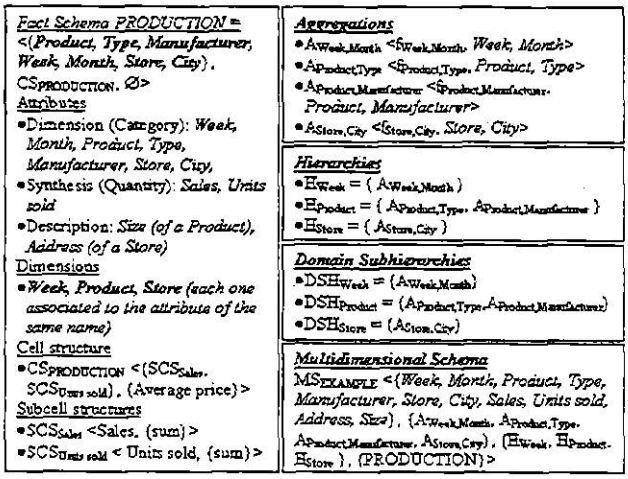

Figura 4. Ejemplo de esquema multidimensional en IDEA

5. Conclusiones

El desarrollo de un Almacén de Datos se ha convertido en un factor crítico de éxito para muchas compañías. Por tanto no es razonable manejar el proceso de construcción fuera del marco de trabajo de una metodología. Esta no es una aproximación totalmente nueva, pero existen algunas aportaciones en el trabajo llevado a cabo en el proyecto EINSTEIN, como:

- Comprobación y construcción de esquemas analíticos partiendo de los requisitos de los usuarios de datos analíticos, utilizando los esquemas operacionales existentes como un soporte para la creación del esquema conceptual multidimensional.
- Ingeniería Inversa para la obtención de esquema en IDEA a partir de las bases de datos multidimensionales específicas existentes.
- Creación o modificación de esquemas operacionales existentes como resultado de las necesidades de los usuarios analíticos.

Se ha desarrollado una metodología multidimensional general basada en una propuesta para la versión 3 de la Metodología Pública Española de Planificación y Desarrollo de Sistemas de Información (METRICA) [9]. Parte de la metodología está soportado por una herramienta CASE (IDEA-DWCASE), de la cual ya se dispone de un primer prototipo presentado en [10]. Esta herramienta, cuya estructura del repositorio es el metamodelo de IDEA, permite la creación de esquemas conceptuales multidimensionales basados en IDEA, y la traducción de éstos a diferentes esquemas lógicos directamente soportados por productos MOLAP o ROLAP (en este momento, la herramienta CASE traduce esquemas IDEA a EXPRESS™ y ORACLE™).

Agradecimientos

Este trabajo se está llevando a cabo dentro del proyecto MIDAS, parcialmente financiado por la CICYT y la comunidad económica europea (2FD97 2163)

6. Referencias

[1] L. Cabibbo y R. Torlone. "A Logical Approach to Multidimensional Databases" En Sixth International Conference on Extending Database Technology (EDBT'98), Valencia, España, Lecture Notes in Computer Science 1377, Springer-Verlag, pages 183-197, 1998.

[2] S. Chaudhuri y U. Dayal. "An Overview of Data Warehousing and OLAP Technology". ACM SIGMOD Record 26(1) Marzo 1997

[3] B. Dinter, C. Sapia, M. Blaschka, G. Höfling. "OLAP Market and Research: Initiating the Cooperation". Journal of Computer Science and Information Management, Vol 2, N. 3, 1999.

[4] S. Gatziu, M. A. Jeusfeld, M. Staudt y Y. Vassiliou. "Design and Management of Data Warehouses - Report on the DMDW'99 Workshop". SIGMOD Record 28(4), Dic. 1999.

[5] Golfarelli, M., Maio, D. and Rizzi, S., "Conceptual design of data warehouses from E/R schemes" en: 31st Hawaii Internaional Conference On System Sciences, 1998.

[6] M. Golfarelli Y S. Rizzi "Designing The Data Warehouse: Key Steps And Crucial Issues". Journal Of Computer Science And Information Management, Vol 2, N. 3, 1999.

[7] R. Kimball. The Data Warehouse Toolkit: Practical techniques for building dimensional data warehouses. John Wiley & Sons, 1996

[8] R. Kimball, L. Reeves, M. Ross, W. y Thornthwaite. The Data Warehouse Lifecicle Toolkit., John Wiley & Sons, Inc., 1998

[9] A. de Miguel et al., "METRICA Version 3: Planning and Development Methodology of Information Systems. Designing a Methodology: A practical experience", in CIICC98, Aguascalientes, México, Nov. 1998. Págs 264-276

[10] A. de Miguel et al. "IDEA-DWCASE: Modeling mutidimensional databases" EDBT 2000 Software Demonstrations track. Konstanz, Alemania, Marzo de 2000

[11] A. Sánchez, J.M. Cavero y A de Miguel. "IDEA: A conceptual multidimensional data model and some methodological implications". CIICC'99, Cancún, Méjico. Págs 307-318.

[12] C. Sapia, G. Höfling, M. Müller, C. Hausdorf, H. Stoyan y U. Grimmer. "On Supporting the Data Warehouse Design by Data Mining Techniques" To appear in GI-Workshop: Data Mining und Data

Warehousing, September 27.-28., 1999, Magdeburg, Germany.

[13] J. Srivastava y P-Y. Chen. "Warehouse Creation - A Potential Roadblock to Data Warehousing". IEEE Transactions on Knowledge and Data Engineering. Vol 11, Num. 1, Ene/Feb 1999

[14] P. Vassiliadis. "Gulliver in the land of data warehousing: practical experiences and observations of a researcher". Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000), Estocolmo, Suecia, Junio-2000

[15] P. Vassiliadis y T. Sellis. "A Survey of Logical Models for OLAP Databases" Sigmod Record, Vol. 28 Num. 4 Dic. 1999

[16] M.C. Wu y A.P. Buchmann, "Research Issues in Data Warehousing". BTW'97, Ulm, March, 1997

[17] A. C. Stylianou y R. L. Kumar, "An integrative framework for IS quality management", Communications of the ACM, Sep. 2000, Vol. 43, N° 9