Metodología para el Diseño Conceptual de Almacenes de Datos

Leopoldo Zepeda[1], Matilde Celma[2]
[1]Departamento de Sistemas y Computación
Instituto Tecnológico de Culiacán
Juan de Dios Bátiz s/n, Col. Guadalupe, 80220
Culiacán, Sinaloa, México
lzepeda@dsic.upv.es
[2]Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022
Valencia, España
mcelma@dsic.upv.es

RESUMEN

El construir un Almacén de Datos (AD), requiere técnicas de diseño diferentes a las usadas en los sistemas tradicionales. En este artículo se presenta una propuesta para el diseño conceptual de almacenes de datos en base a dos perspectivas, las bases de datos operacionales existentes y los requisitos de usuario.

Palabras claves: Almacén de datos, Diseño conceptual, esquema operacional, requerimientos de usuario

1. INTRODUCCIÓN

Un AD es una base de datos diseñada para la toma de decisiones, donde se integra una gran cantidad de información histórica proveniente de diferentes fuentes de datos y contiene información variante en el tiempo más no volátil [1]. Estas características especiales, han abierto nuevas oportunidades de estudio en el área de bases de datos: Optimización de consultas, técnicas de indexación y herramientas para la explotación de datos orientadas al usuario final, sin embargo se ha mostrando poco interés en los aspectos relacionados al diseño conceptual lo cual se debe a que la tecnología AD se originó en el ámbito industrial, donde por lo general no se le da importancia a los aspectos conceptuales.[2]

Los nuevos objetivos de esta tecnología son la definición de modelos de diseño de datos, en términos generales las propuestas existentes siguen un modelo de datos multidimensional, el cual mas que un modelo es una filosofía de modelado heredada de las herramientas para el análisis de datos (OLAP[1]). En un esquema multidimensional la actividad de la organización objeto de análisis (hecho) es modelada.

El esquema multidimensional representa esta actividad en el centro, con sus indicadores mas relevantes (medidas), y las dimensiones alrededor de el, donde cada dimensión es descrita como un conjunto de atributos [3].

La representación grafica multidimensional puede ser variada cubos n-dimensionales donde cada eje representa una dimensión y las celdas representan hechos acerca de la actividad o un esquema estrella, donde en el medio de la estrella están los hechos y alrededor las dimensiones de análisis. En términos generales, los modelos propuestos, incorporan conceptos o constructores presentados en los modelos de datos aceptados universalmente. Esas propuestas son presentadas como extensiones a los modelos clásicos o como nuevos modelos los cuales resaltan el aspecto multidimensional de los datos.[4,5,6]

Podemos concluir que en esos modelos los conceptos clásicos son adaptados al modelado de un AD, por que un esquema conceptual de un AD y el esquema conceptual de un operacional tienen estructuralmente pocas diferencias aunque tienen diferentes medios de explotación y diferentes consultas.

2. TRABAJOS RELACIONADOS

Las investigaciones sobre la tecnología de AD y OLAP pueden encontrase en diversos libros y artículos de interés, uno de los libros mas importantes es [3], entre las investigaciones más importante se encuentra el proyecto realizado en la Universidad de Stanford, el cual se enfoca al desarrollo de algoritmos relacionados con la integración y mantenimiento de la información extraída a partir de fuentes de datos heterogéneas y autónomas [7]. En general se encuentran pocas contribuciones en la literatura que conciernan específicamente con el diseño conceptual de AD, a continuación se muestran las más relevantes:

[1] Del ingles "on-line analytical processing"

J. Díaz de León. G. González, J. Figueroa (Eds.): Avances en Ciencias de la Computación, pp. 272-276, 2003.
© IPN, México 2003.

Cabilbbo & Ricardo Torlone [8] de la universidad de Roma, ilustran el método de diseño Md, el cual es un modelo lógico para sistemas OLAP y muestran como puede ser usado para el diseño de bases de datos multidimensionales. El método de diseño propuesto construye un esquema Md iniciando de una base de datos operacional existente, donde el esquema Md consiste de un conjunto finito de dimensiones y un conjunto finito de F-Tables. Los esquemas Md se obtienen a través de cuatro pasos: Identificación de hechos y dimensiones, reestructuración del diagrama entidad relación, derivación de un grafo dimensional y finalmente el traslado a un modelo Md.

Golfarelly M. & Dario M [9,10] de la universidad de Bologna, Propone un método para la transformación de diagramas entidad relación a una estructura de árbol. Este modelo representa la tabla de hechos como la raiz del árbol y los atributos de dimensiones como el árbol descendente. Los elementos del árbol son los hechos, atributos, dimensiones y jerarquías; otros elementos los cuales son representados son la aditividad de los atributos de hechos con las dimensiones.

3. METODOLOGÍA DE SOLUCIÓN

El problema de diseño de un Almacén de Datos consiste en obtener un conjunto de esquemas multidimensionales que permita capturar los requisitos de usuario y que puedan ser mantenidos por las bases de datos operacionales existentes.

El objetivo del trabajo consiste en definir una metodología que permita realizar el diseño basándose en los supuestos anteriores, para abordar el problema se puede dividir en tres fases: 1) Obtener un conjunto de esquemas multidimensionales a partir de los diagramas E/R, 2) Obtener un conjunto de métricas que permitan seleccionar un esquema multidimensional considerando las consultas de usuario. 3) Refinamiento manual del esquema obtenido. A continuación se explica cada una de las fases:

3.1. Fase 1: *Obtener un conjunto de esquemas multidimensionales a partir de los diagramas E/R*

La mayoría de los sistemas de información actuales son implementados sobre tecnología relacional. En la construcción de estos sistemas el diseño de la base de datos ocupa un lugar relevante. Para realizar el diseño las metodologías existentes plantean tres fases: diseño conceptual, diseño lógico y diseño físico. Para el diseño conceptual se utiliza distintos modelos siendo uno de los más extendidos el modelo E/R.

Esta metodología de fases es extensible al diseño de AD, pero existe una diferencia fundamental: el AD debe ser mantenido a partir del sistema operacional existente. Esta reflexión nos conduce a plantear el diseño conceptual a partir del esquema conceptual del sistema operacional y de los requisitos de usuario. La metodología para obtener un conjunto de esquemas multidimensionales a partir de un diagrama E/R, consiste en realizar un análisis exhaustivo del diagrama E/R con el fin de identificar las entidades que son candidatas a ser hechos, una vez identificadas las entidades de hechos se realiza una búsqueda a partir de cada hecho identificado con el fin de agregar dimensiones, produciendo un esquema multidimensional para cada entidad de hechos candidata.

Para el funcionamiento correcto de esta metodología se asume lo siguiente:

- Las relaciones N-ary (N>2) o las relaciones binarias se han convertido en relaciones Uno-Muchos, creando nuevas entidades.
- La generalización se ha convertido en relaciones Uno a Uno.

Una vez que se dispone del diagrama E/R con esas caracteristicas, se deben seguir los siguientes pasos: Identificar los hechos, definir las dimensiones, definir las jerarquías en cada dimensión.

3.1.1. Identificar los hechos

Esta actividad consiste en realizar un análisis detallado del diagrama E/R con el objetivo de seleccionar las entidades que son candidatas a ser hechos en el esquema multidimensional. En el contexto de un diagrama E/R, una entidad H se clasificará como entidad de hechos si cumple con una de las siguientes caracteristicas: a) H contiene al menos un atributo numérico no primario (no forma parte de la llave primaria) y se encuentra relacionada al menos con una entidad E_1, con cardinalidad Muchos a Uno, b) H no contiene atributos y se encuentra relacionada con otras entidades con cardinalidad Muchos a Uno.

Debido a que no todas las entidades identificadas como entidades de hechos son de interés para la toma de decisiones, hasta que los requisitos de usuario sean identificados las entidades en esta etapa son candidatas a ser entidades de hechos en el esquema multidimensional final y se almacenarán en una lista de entidades F.

3.1.2. Definir las dimensiones y jerarquías

Una dimensión es un subesquema de un esquema E/R, que representa un punto de vista desde el cual

el análisis de un hecho se puede realizar. Durante este proceso es necesario considerar lo siguiente:

- Todas las entidades del diagrama E/R que no son candidatas a ser entidades de hechos son dimensiones candidatas.
- Una dimensión puede conectarse con más de una entidad de hechos.
- Cuando exista una relación entre entidades identificadas como hechos, la entidad hija heredará todas las dimensiones de la entidad padre.
- La dimensión Tiempo siempre debe formar parte del esquema multidimensional, por lo que se debe añadir al esquema multidimensional.

El resultado de esta fase es un conjunto de esquemas multidimensionales (Uno para cada entidad de hechos identificada), que serán candidatos a formar parte del esquema final del AD, este conjunto de diagramas puede obtenerse de manera automática al aplicar el algoritmo de la Fig. 1, que funciona de la siguiente manera: El algoritmo parte de la lista de entidades de hechos candidatas F y explora todas las rutas posibles para encontrar dimensiones y jerarquías. El algoritmo recibe como entrada el diagrama E/R y F, la salida es una lista D de esquemas multidimensionales para cada entidad de F, (donde cada entidad de F es el centro del esquema multidimensional). El algoritmo considera al diagrama E/R como un grafo y añade a cada entidad del conjunto F el atributo "visitado", e inicia un recorrido en profundidad a partir de cada una de ellas las cuales actuaran como nodo raíz del recorrido, la búsqueda de las dimensiones y jerarquías se realiza de la siguiente forma:

Comienza el recorrido en la entidad de hechos candidata E_I, considerada como raíz del recorrido y la marca como visitada. Elige una entidad E_k adyacente a E_I, por medio de una relación Muchos a Uno (a la que llamaremos en el algoritmo Fk_J).

```
F = (Conjunto de entidades hechos candidatas)

Para Cada E_i ∈ F hacer
D[E_i]=( )
Búsqueda_dim(E_i)
Fin para

Procedimiento Búsqueda_dim(E_i :Entidad ∈ ER )
Inicio
If E_i ∈ F entonces E_i.visitado=cierto
For each Fk_j ∈ E_i
        Ek = Entidad relacionada a través de Fk_j
        If Ek.visitado=cierto Then
                D[E_i]= D[E_i] U D[Ek]
        Else
        If not Ek ∈ F then
                D[Ei]=D[Ei] U (Ek)
        Búsqueda_Dim(Ek)
Fin For
Fin de procedimiento
```

Fig. 1. Algoritmo para obtener las dimensiones y jerarquías.

La parte principal del algoritmo es el procedimiento recursivo Búsqueda_dim, que inicia marcando las entidades E_I que son consideradas hechos candidatos como visitada, a partir de las relaciones Muchos a Uno de E_I, asocia una entidad E_k que formara parte del diagrama multidimensional asociado a E_I, el diagrama multidimensional para cada hecho candidato E_I es almacenado en la lista D.

Para explicar el algoritmo utilizaremos el esquema de la Fig. 2. El conjunto de hechos F identificados en el diagrama es: F = {Artículo, Línea, Tienen, Ticket}. En la primera iteración del ciclo se selecciona el elemento E_I del lista F y se le asocia una posición en el lista D, por ejemplo E_1= *Artículo*, D[Artículo]={}. El procedimiento recursivo Busqueda_Dim(E_I), recibe como argumento de entrada *Artículo*, y debido a que se encuentra en la lista de hechos F es marcada como visitado.

El procedimiento realiza una búsqueda en el diagrama de todas las entidades relacionadas con E_1 (*Artículo*), las cuales serán consideradas niveles de dimensión de E_1, y serán almacenadas en la lista D. Una vez finalizado el procedimiento recursivo Búsqueda_dim(),se sigue con la búsqueda de dimensiones para el segundo elemento de F. Así E_2= Línea, y se inicia un recorrido por cada Relación Muchos a Uno de ella. En el diagrama de la Fig. 2, se muestran las entidades con las que se relaciona la entidad Línea.

Debido a que existe una relación entre Línea y Artículo, y que Artículo esta marcado como visitado las entidades de D[Artículo], pasan a formar parte de D[Línea]. Por lo que D[Línea]=D[Línea] U D[Artículo], el esquema multidimensional para Línea se muestra en el Fig. 3.

Esta fase devuelve todos los posibles esquemas multidimensionales que pueden extraerse del diagrama E/R del operacional, en la Fig. 4, se muestra el conjunto que deberán ser refinados a partir de los requerimientos de usuario.

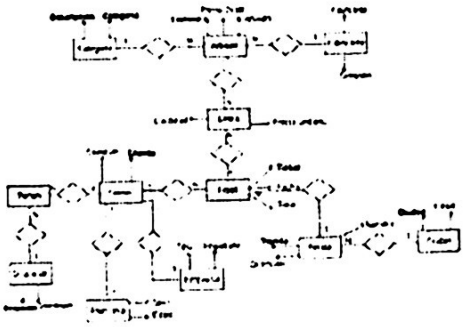

Fig. 2. Diagrama con los hechos candidatos.

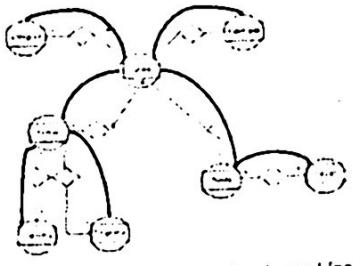

Fig. 3. Esquema multidimensional para Línea.

3.2. Fase 2: ***Obtener un conjunto de métricas que permitan seleccionar los esquemas multidimensionales que soportan las consultas de usuario.***

Para seleccionar un esquema multidimensional es necesario contrastar los esquemas multidimensionales con las consultas que plantea el usuario en el sistema operacional. Esta comparación se basa en el número de elementos de los esquemas multidimensionales que aparecen directamente en las consultas. Nosotros requerimos principalmente una correspondencia obligatoria entre los hechos que son el centro del esquema multidimensional y una tabla de la cláusula FROM.

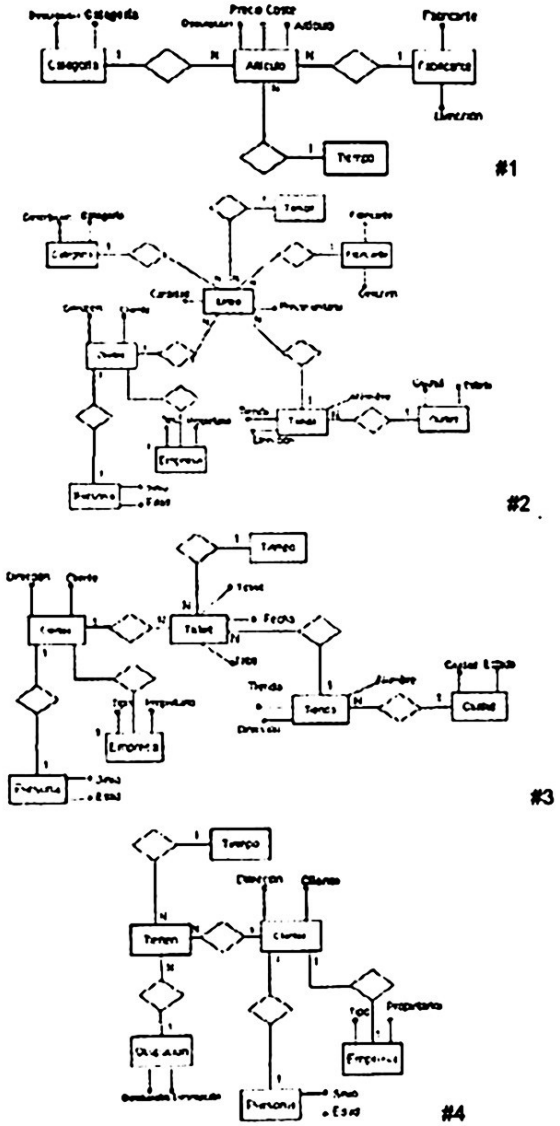

Fig. 4. Conjunto de esquemas multidimensionales.

Las métricas utilizadas para decidir si un esquema multidimensional captura los requerimientos de usuario son las siguientes:

Correspondencia de atributos. Se debe contar el número de atributos de la tabla de hechos que se corresponden directamente con los atributos de la cláusula SELECT.

- *Correspondencia de dimensiones. Se debe contar el número de dimensiones del esquema multidimensional que se corresponden directamente con las tablas de la cláusula FROM.*
- *No correspondencia de atributos.* Se debe contar el número de atributos de la tabla de hechos que no se corresponden directamente con los atributos de la cláusula SELECT.
- *No correspondencia de dimensiones.* Se debe contar el número de dimensiones del esquema multidimensional que no se corresponden directamente con las tablas de la cláusula FROM.
- *Dimensiones que Faltan.* Se debe indicar que dimensión falta en el esquema multidimensional para dar respuesta a la consulta de usuario.

Hay dos aspectos de una consulta que son usados para determinar si un esquema candidato puede responder a una consulta: las tablas en la cláusula FROM y los atributos numéricos en la cláusula SELECT. Si un esquema candidato no contiene una entidad de hechos que se corresponda con la tabla que en la cláusula FROM contiene los atributos numéricos requeridos en la consulta éste no puede dar respuesta a la consulta, de lo contrario será necesario verificar con cuantas entidades de dimensiones se corresponden las otras tablas de la cláusula FROM y con cuantas no se corresponden. Para mostrar cómo funciona esta fase, considere las consultas Q1 y Q2.

Q1: Muestra los 10 artículos que tuvieron menos ventas durante el año 2000.

```
SELECT TOP 10 SUM(Cantidad) AS Total, Articulo.descripción
FROM Linea, Articulo, Ticket
WHERE   Linea.id_articulo=Articulo.id_articulo AND
        Linea.Ticket_id= Ticket.Ticket_id AND Datepart(YY,Ticket.Fecha)=2000
GROUP BY Articulo.descripción
ORDER BY SUM(Cantidad) DESC
```

Q2: Muestra las cinco categorías de artículos que tuvieron mayor promedio de ventas durante el año 2000.

```
SELECT TOP 5 AVG(Precio Unitario*Cantidad) AS Total, Categoria.descripción
FROM Linea, Articulo, Categoria, Ticket
WHERE Linea.id_articulo=Articulo.id_articulo AND
        Articulo.id_categoria=Categoria.id_categoria AND
        Linea.id_linea=Ticket.id_linea and datepart(YY,Ticket.Fecha)=2000
GROUP BY Categoria.descripción
ORDER BY AVG(Total)
```

La información registrada en la tabla de la Fig. 5, refleja que el esquema #2 da respuesta a las dos consultas, también se muestra el número de dimensiones con las que tienen relación la consulta,

pero también se muestra que la dimensión Artículo que es requerida por las consultas no es parte del esquema multidimensional #2.

Consulta	Q1	Q2
Esquema Multidimensional	#2	#2
Correspondencia de atributos	1	2
Correspondencia de dimensiones	0	1
No Correspondencia de atributos	1	0
No correspondencia de dimensiones	2	3
Dimensiones que falta	Ticket, Artículo	Ticket, Artículo

Fig. 5. Relación de consultas y esquemas.

A partir de la información registrada en la tabla de correspondencia se puede deducir que el esquema multidimensional #2 es el que satisface las consultas pero también observamos que no incluye la dimensión Artículos y que esta dimensión aparece como un hecho en el esquema #1. Por lo que debe de realizarse un refinamiento del esquema multidimensional y agregar como parte del esquema #2 el esquema multidimensional #1, Fig. 6.

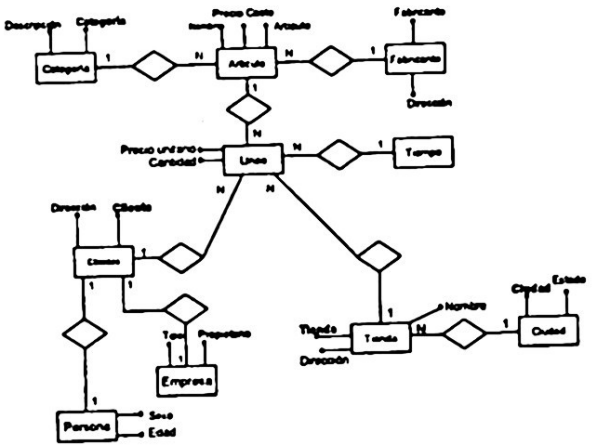

Fig. 6. Esquema multidimensional final.

4. CONCLUSIONES

En este trabajo se presenta una propuesta para el modelado conceptual de almacenes de datos. La metodología se basa en el uso del modelo E/R y se apoya en el diagrama E/R del sistema operacional y en los requisitos de usuario. Para ello se ha desarrollado una metodología de tres fases.

La primera fase se inicia con la identificación de las entidades de hechos basándose en un conjunto de criterios definidos. A continuación, partiendo de estas entidades de hechos y del esquema E/R del operacional se obtiene un conjunto de esquemas multidimensionales, para ello se proporciona un algoritmo recursivo. En la segunda fase se obtiene un conjunto de métricas a partir de las consultas que el usuario plantea con más frecuencia al sistema operacional, estas métricas permitirán seleccionar un esquema multidimensional del conjunto de esquemas candidatos. En la tercera fase se hace un refinamiento manual a partir de los requisitos de usuario, con el objetivo de representar las restricciones especiales del esquema multidimensional.

Como trabajos futuros se pretende extender esta propuesta, en particular mejorar la fase 2. Se espera ampliar el conjunto de métricas que permiten seleccionar el esquema multidimensional entre todos los esquemas candidatos, para ello será preciso definir o adoptar una metodología que permita obtener y analizar los requisitos de los potenciales usuarios del AD.

Referencias

[1] Inmon, W.H. "Building the Data Warehouse", ed. Jhon Wiley and Sons, Inc. New York, NY. 1996

[2] Abello, Samos y Saltor, "YAM2: A Multidimensional Conceptual Model", tesis doctoral, Universidad Politécnica de Cataluña, 2002.

[3] Kimball, R. "The Data Warehouse Tollkit: Practical Techniques for building Dimensional Data Warehouses", John Wiley and Sons, Inc., New York, NY, 1998.

[4] Ashish Gupta, Inderpal Singh M. "Maintenance of materialized View: Problems, Techniques, and aplications", 1999

[5] Husemann, Lectenborger, and Vosen, "Conceptual Data Warehouse Design", Proceeding of the Intl. Workshop on DMDW 2000.

[6] Carsten S., "Extending the E/R Model for the Multidimensional Paradigm, www.citeseer.nj.nec.com/434096.html.

[7] J.Widom, "Research problem in data warehousing", www.dbpubs.stanford.edu/pub/1995-24

[8] Cabbibo, L and Torlone, R. "A Logical Approach to Multidimensional Databases", www.citeseer.nj.nec.com/cabibbo98logical.html. 1998.

[9] Golfarelli, M., Maio, D. Rizzi, S.,"Conceptul Design of Data Warehouse from E/R Schemas", Proc. 32th HICSS 1998.

[10] Golfarello, D. Rizzi, "A methodological framework for data warehouse Design", Dolap 1999.