

RESUMEN

BASES DE DATOS : OPTIMIZACION DEL MODELO INTERNO

Rodolfo Schmal S (★)
Yanko Ossandón N.(★)

Cuando se está ante la tarea de diseñar una base de datos -BD- debe abordarse la tarea de captar toda la semántica contenida en la realidad que interesa recoger, y por otro lado debe procurarse que el rendimiento sea aceptable. Por lo general la atención ha estado centrada en el primer objetivo, descuidándose el segundo de ellos. Al tratar la temática del rendimiento estamos haciendo referencia al tiempo de respuesta que involucre todo acceso que se desee realizar sobre la BD. Una BD puede satisfacer las exigencias lógicas propias de la realidad percibida, pero no será implementable si a la hora de hacer uso de ella los tiempos de respuestas son excesivos. Este trabajo presenta una metodología destinada justamente a optimizar la representación interna considerando las actividades que se estiman se van a realizar sobre la BD.

Arica, Julio 1987

(★)Masters en Informática Universidad Politécnica de Madrid
Académicos del Departamento de Computación e Informática
Universidad de Tarapacá - Arica - Chile

BASES DE DATOS : OPTIMIZACION DEL MODELO INTERNO

Rodolfo Schmal S (★)
Yanko Ossandón N.(★)

1.0 INTRODUCCION

El diseño de una base de datos (BD) es una tarea compleja, que puede tomar meses de trabajo, incluso años. Hasta ahora, el diseño ha sido un arte más que una ciencia, carente de estándares y normas de procedimiento, sujeto a la intuición y experiencia de sus diseñadores.

Determinar la semántica de los datos es uno de los mayores obstáculos a enfrentar, y una vez recogida esta semántica, las estructuras diseñadas pueden ser tan complejas que resulte impracticable su operación, dado el imperativo de eficiencia que exige su implementación. En este sentido la BD debe ser capaz de proporcionar la información requerida con una eficiencia razonable. Un diseño simplista puede dificultar la programación al necesitar una lógica más complicada para acceder a la información, influyendo además en un mayor tiempo y costo de implementación, y posteriormente en mayor tiempo de proceso.

Aspectos como la integridad y la consistencia de los datos, la privacidad y la seguridad, amén de otras consideraciones, son las que nos invitan a buscar diseños sistemáticos que nos pueden ayudar a reconocer y resolver muchas de estas exigencias, antes que la BD sea instalada.

Una de las consecuencias más graves que ocasiona un diseño deficiente es el costo prohibitivo que implica reestructurar una base ineficiente.

Generalmente, las mayores dificultades que se presentan las podemos agrupar en las siguientes categorías :

- Análisis insuficiente de las necesidades de los usuarios.
- Estructuración más allá de la capacidad humana para su análisis manual.
- Dificultades para predecir la incidencia que la implantación del sistema ocasionará en el medio ambiente existente (coexistencia con otras BD).

(★)Masters en Informática Universidad Politécnica de Madrid
Académicos del Departamento de Computación e Informática
Universidad de Tarapacá - Arica - Chile

- Exigencias de cumplimientos de plazos por parte de la administración.

La búsqueda de metodologías de diseño sistemático, y de técnicas de diseño asistido por computador han significado importantes contribuciones para minimizar estos problemas, además de mejorar la calidad y acortar el tiempo de diseño.

En este contexto utilizaremos el formalismo individual para modelar la realidad que nos interesa automatizar, pero colocando énfasis en la optimización del modelo interno lógico.

El formalismo individual es una metodología propuesta por un grupo de trabajo de la Universidad de Marsella (H. Tardieu, D. Nanci, D. Pascot [1]) que ha tenido una amplia acogida en Francia y Europa, que tiene sus raíces en el modelo entidad-relación de P.P.Chen[4].

2.0 FORMALISMO INDIVIDUAL

Nuestro objetivo básico es representar aspectos del mundo real que nos interesa estudiar, en un medio susceptible de ser manipulado electrónicamente. Para llegar, desde los objetos de ese mundo real, a este otro de los datos es preciso realizar una importante labor de abstracción y estructuración que se ha de apoyar en un modelo que nos proporcione un conjunto de métodos, reglas y símbolos.

El formalismo individual permite definir y describir el real percibido utilizando tres elementos [1]:

- los INDIVIDUOS-TIPO que son los objetos percibidos, concretos o abstractos, equivalentes a las ENTIDADES del modelo entidad-relación[2];
- las RELACIONES-TIPO que son las asociaciones entre los objetos percibidos;
- las PROPIEDADES que son las características que permiten describir a los individuos y sus relaciones.

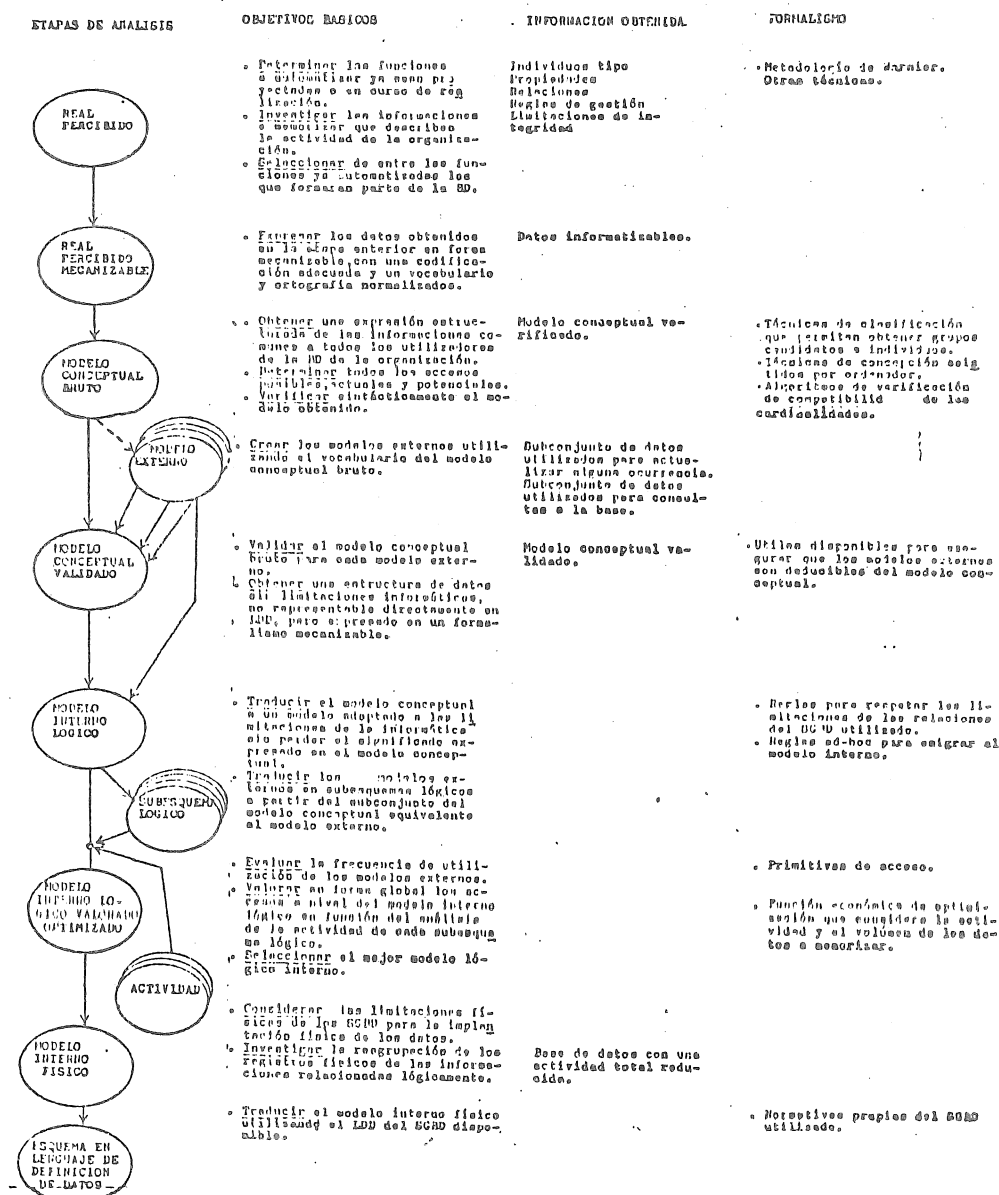
Nuestro trabajo, como informáticos, consistirá en acotar el espacio de información al subespacio que corresponde al fenómeno que se está estudiando con el objeto de conseguir minimizar el tiempo de acceso a los datos y reducir al máximo el soporte físico necesario para el almacenamiento. Normalmente estos dos objetivos se contraponen entre sí, por lo que suele tener que alcanzarse alguna solución de compromiso.

Una de las dificultades claves reside en describir directamente el real percibido en el lenguaje de descripción del SGBD que se vaya a utilizar; por ello es necesario superar las siguientes etapas:

1. ELABORAR EL MODELO CONCEPTUAL.
2. VALIDAR EL MODELO CONCEPTUAL.
3. TRANSFORMAR EL MODELO CONCEPTUAL EN UN MODELO INTERNO LOGICO
4. TRANSFORMAR EL MODELO INTERNO LOGICO EN UN MODELO INTERNO FISICO.

El significado de los términos modelo conceptual, modelo interno lógico y modelo interno físico es el dado en [1].

Una síntesis de ellas se presenta en la siguiente figura :



De la metodología propuesta nos centraremos en el tema que denominamos optimización del modelo interno.

3.0 OPTIMIZACION DEL MODELO INTERNO

3.1 ESTUDIO DE LA ACTIVIDAD DE LOS MODELOS EXTERNOS

Los programas que actúan sobre una BD están destinados a satisfacer las necesidades de los usuarios expresadas mediante modelos externos -subesquemas en el modelo Codasyl-. En la gran mayoría de los casos las acciones que con ellos se realizan apuntan a crear, eliminar, consultar y modificar datos. Su campo de acción puede ser muy variado y afectar a diferentes volúmenes de datos.

Para poder evaluar esta actividad -entendiendo por tal al producto de acciones elementales por la frecuencia de utilización del programa-, usaremos un conjunto de primitivas de acceso a la BD, que representan funciones específicas, y que definiremos sobre el modelo interno lógico[1].

Algunas de estas funciones son :

1. Accesar una ocurrencia de un registro
2. Incorporar una ocurrencia de un registro
3. Suprimir una ocurrencia de un registro
4. Modificar una ocurrencia de un registro

Dado un esquema de BD, la acciones a realizar para poblar la base de datos se expresan en términos de primitivas valorizadas considerando las cardinalidades medias de cada set y la cantidad de registros nuevos a incorporar. Para obtener una valoración global de los accesos a nivel de modelo interno logico, deberá analizarse y valorar cada uno de los subesquemas lógicos e ir acumulando el conjunto de las actividades. De esta manera se obtiene un modelo interno lógico valorado.

El modelo interno obtenido es una traducción del modelo conceptual a algún modelo implementable, en este caso Codasyl, sin considerar el rendimiento del modelo; por lo tanto debemos tratar de minimizar una función económica (FE) que nos permitirá seleccionar el mejor modelo interno.

Una expresión para esta función FE podría ser:

$$FE = (NA \times PA) + (NC \times PAL)$$

donde

NA = número de accesos por unidad de tiempo
 PA = precio del acceso por unidad de tiempo
 NC = número de caracteres almacenados
 PAL = precio del almacenamiento de un carácter por unidad de tiempo

La utilización de esta función supone considerar las siguientes hipótesis :

- . La actividad de una aplicación puede medirse en número de accesos, independientemente de la programación de dicha aplicación;
- . La medida determinada no considera los efectos borde, entendiéndose como tales a los accesos extras producidos por el uso de pistas de desbordamiento o por falta de reorganización de los datos;
- . El volumen de datos accesorios (punteros) necesarios en la implantación física de la BD es constante, cualquiera que sea la técnica utilizada; y
- . Las máquinas empleadas tienen jerarquía de memoria simple: principal y secundaria; el volumen almacenado y el número de accesos a la memoria secundaria son los factores antagónicos para los cuales es necesario buscar un equilibrio óptimo.

Por otra parte, y a nivel de modelo interno, se supone :

- * que los volúmenes almacenados provienen de los campos contenidos en los registros; y
- * que los accesos a la memoria secundaria provienen únicamente de la actividad definida por los ME sobre el modelo interno en término de primitivas.

3.2 VALORACION DE LAS PRIMITIVAS EN TERMINOS DE ACCESOS

El objetivo de esta etapa es obtener una medida uniforme de la actividad del modelo interno lógico.

A cada primitiva de acceso se le asigna un número equivalente, en términos de acceso, y que sea compatible con los accesos del sistema de gestión que se utiliza [5].

3.3 PROCEDIMIENTOS DE OPTIMIZACION

Para optimizar el modelo podemos utilizar varios criterios :

Crear redundancia

- . Repetir algunas informaciones con el fin de disminuir la actividad;
- . Hacer emigrar algunos campos que provocan accesos continuos a los registros cabecera, desde un registro cabecera hacia un registro miembro;
- . Crear sets redundantes para evitar ciertos recorridos; esto corresponde a crear sets equivalentes a la composición de dos o más sets que forman una jerarquía. De esta manera se economizan accesos, especialmente en las operaciones de acceso a la cabecera.

Abandonar algunos sets ordenados

- . A la luz de los casos analizados, está claro que mantener sets ordenados implica realizar un considerable número de accesos cuando se efectúa una actualización. Si se elimina el ordenamiento de algunos sets de cardinalidad media muy elevada, se disminuye notablemente la actividad. Debe considerarse además que esta optimización se hace sin incrementar el almacenamiento.

Revisar el modelo conceptual

- . Otra posibilidad consiste en reestudiar la definición de aquellas entidades que se utilizan en forma disjunta por dos subesquemas lógicos de muy elevada tasa de actividad, separándolas en dos entidades, lo que nos obligará a definir una relación adicional, si algún otro subesquema utiliza la entidad original, para indicar que las dos nuevas entidades están a su vez relacionadas en forma biunívoca.

Revisar las cardinalidades

- . La posibilidad de definir en el modelo conceptual relaciones n-arias o binarias del tipo 0-n - 0-n, es una riqueza semántica de la cual puede abusarse. Dimensionar adecuadamente el modelo reducirá considerablemente la actividad del modelo interno.

Conociendo la valoración del modelo interno lógico, se reestructuran los registros y sets en base a los criterios expuestos anteriormente, redefiniendo los subesquemas que correspondan y valorizando el nuevo modelo interno resultante.

El modelo final será aquel que minimice la función económica y que mantenga un equilibrio entre el número de accesos y el volumen de datos almacenados.

Este es un proceso tedioso y sólo será eficaz en la medida que se disponga de un producto utilitario que permita obtener y valorar subesquemas lógicos.

Al término de esta fase de optimización se podría afirmar que el diseñador ha alcanzado un compromiso (trade-off) entre volumen y acceso, independiente del SGBD a usar.

4.0 CONCLUSIONES

En múltiples ocasiones hemos estado frente a BD que satisfacen en plenitud las exigencias lógicas de los usuarios en cuanto a la definición de los individuos, sus asociaciones, y sus propiedades para que recojan la semántica incorporada en la realidad a percibir; sin embargo a la hora de la implementación la problemática del rendimiento pasa a ser crucial. Esto último obliga a buscar modelos internos que maximicen el rendimiento de una BD dada. Este trabajo se inscribe justamente en la obtención de alguna metodología que nos permita determinar las estructuras internas mas adecuadas para una realidad dada. Esta última dada no solo por los elementos que la componen, sus características, y sus relaciones, sino incorporando como parte de esa realidad lo que en el presente trabajo hemos llamado el nivel de actividad de sus elementos.

Queda pendiente lo relativo a la automatización de la metodología aquí sugerida, un afinamiento de ella, su extensión a sistemas de gestión de BD basados en modelos distintos al CODASYL, y su aplicación fuera de un marco académico.

Entre las limitaciones que presenta lo sugerido no podemos ignorar la exigencia que información que no siempre está disponible cuando se aborda la tarea de diseñar una BD. El conocimiento de las cardinalidades medias de los individuos y sus asociaciones, es clave en la aplicación de esta metodología. También lo es el conocimiento del comportamiento futuro de estas cardinalidades. Prever el comportamiento que tendrán los elementos constitutivos de la BD nos ahorrará futuras reorganizaciones altamente indeseables por lo que conlleva en consumo de recursos de toda índole. Otra limitación que debemos mencionar radica en su carácter determinístico. Una alternativa más acuciosa nos debe conducir al manejo de modelos probabilísticos. No obstante lo anterior hemos optado por un conocimiento determinístico de las variables comprometidas en aras de su simplicidad y facilidad de implementación.

5.0 BIBLIOGRAFIA

- [1] CONCEPTION D'UN SYSTEME D'INFORMATION
H. Tardieu, D. Nancy, D. Pascot
Les Editions d'Organisation, Paris, 1979.
- [2] UNA METODOLOGIA DE DISEÑO LOGICO DE BASES DE DATOS
Rodolfo Schmal S., Tesis de Maestría, 1985.
- [3] COMPUTER-ASSISTED DATA BASE DESIGN
George V. Hubbard
Van Nostrand Reinhold Company, 1981.
- [4] ENTITY-RELATIONSHIP MODEL-TOWARD A UNIFIED VIEW OF DATA
P.P. Chen, ACM TODS, VOL 1 Nro 1, Marzo 1976.
- [5] DISEÑO CONCEPTUAL DE UNA BASE DE DATOS DE SEGUIMIENTO
CURRICULAR Y CONTROL ACADÉMICO
Yanko Ossandón N., Tesis de Maestría, 1984.