Capítulo 3.
Integración de esquemas en MDBMS

Para que el usuario de un MDBMS pueda accesar de manera transparente y uniforme la información almacenada en diferentes componentes de bases de datos, se necesita resolver los conflictos de heterogeneidad semántica y de datos. La idea de resolver los conflictos es lograr una integración de esquemas, que permita a los usuarios de MDBMS formular solo una consulta para *n* Bases de Datos en lugar de *n* consultas, una para cada Base de Datos.

El presente capítulo presenta la clasificación de los conflictos de esquema y de datos; además describe de manera detallada, como es que se presentan los conflictos al integrar las Bases de Datos Componentes (BDC´s). Se revisa y discute la metodología propuesta para lograr la integración de esquemas. La metodología de integración se describe, considerando los conflictos y casos particulares para el integrador de esquemas propuesto en este trabajo de investigación.

3.1 Conflictos de Esquema

Debido a que las BDC´s operan independientemente (sin un control centralizado o coordinador distribuido), éstas pueden presentar discrepancia estructural y de representación. Estas diferencias permiten identificar y clasificar los conflictos de esquema manejados durante el proceso de integración. El conjunto de conflictos identificados por [Kim y Seo, 1991] para la integración de esquemas se describen en los párrafos siguientes.

3.1.1 Conflictos en tablas

Conflictos de nombrado de tablas

a. *Nombres diferentes para tablas equivalentes.*- Ocurre cuando se asignan nombres diferentes a tablas semánticamente equivalentes (sinónimos).

b. *Nombres iguales para tablas no equivalentes.*- Cuando se asigna el mismo nombre a tablas semánticamente diferentes (homónimos).

Conflictos en la estructura de las tablas

a. *Atributos faltantes.*- La conceptualización del diseño de un esquema de BD puede llevar a la omisión de atributos que no se consideren representativos.

b. *Atributos implícitos.*- En ocasiones los atributos existentes pueden ser suficientes para deducir algún otro atributo requerido para la integración.

Conflictos en restricciones de integridad.- En esta categoría se incluyen las deferencias que pueden surgir con respecto a la selección de llaves primarias, secundarias, extranjeras y conflictos en medida de la integridad referencial.

Conflictos en la organización de la información.- El número de tablas requeridas para modelar la BD en cada componente puede diferir de acuerdo a la conceptualización de solución de cada diseñador.

3.1.2 Conflictos en atributos

Conflictos de nombrado de atributos.- El concepto de sinónimo y homonimia aplicado a los conflictos de nombrado de tablas son aplicables a éstos tipos de conflictos.

Conflictos en valores por ausencia.- La definición implícita de algunos valores por ausencia asignados por DBMS, podría llevar a contradicciones en la semántica de los datos.

Conflictos por restricciones de asignación de valores a los atributos

a. *Conflictos en los tipos de datos.*- Los tipos de datos pueden diferir en cuanto al criterio de diseño de cada aplicación.

b. *Conflictos en restricciones de dominio* .- Reglas impuestas para a asignación de valores o consideración en los criterios de unicidad.

Conflictos por la cardinalidad y grado de atomicidad.- El grado de detalle de cada atributo puede ser distinto en cada aplicación. Además, cada modelo de información establece restricciones y posibilidades de estructuración muy diferentes.

Conflictos en la representación de la información.- Puede presentarse el caso de que el mismo concepto se presente como una entidad en una aplicación y en otra como un solo atributo.

3.2 Conflictos de datos

Aún a pesar de tener esquemas de BDC´s equivalentes en cuanto a la estructura de sus tablas y atributos, es posible presentar otra serie de problemas identificados como conflictos de datos, dichos conflictos presentados por [Kim y Seo 1990], se describen a continuación:

Conflictos entre los valores.- Cuando se espera que instancias equivalentes tengan los mismos valores, pero muestran inconsistencia debido a que los datos son capturados incorrectamente o los datos son obsoletos.

Diferencias en la representación.- Situaciones de contexto y cultura organizacional, entre otros factores pueden llevar a que cada BDC seleccione una representación distinta en la información. Dichas inconsistencias pueden presentarse por:

a. *Notaciones diferentes.-* Cuando existen diferentes formas para representar un mismo dato. (p.e. calificaciones numéricas o con escalas de letras)

b. *Unidades distintas.-* La diversidad de unidades, sobre todo para valores numéricos, trae consigo problemas de interpretación. (p.e. la diferencia en el sistema de medición ingles y el internacional)

c. *Diferencias en las representaciones.-* Cuando existen diferentes formas para representar un valor de un atributo. (p.e. para el atributo Estado se puede tener: Tlaxcala, Tlax., Tx, etc.)

3.3 Modelado de datos en los esquemas locales y federados

El modelado de datos es el proceso de crear una representación consistente de los datos del usuarios. Existen diferentes propuestas de modelado, tales como el de red, jerárquico, relacional y orientado a objetos, cada uno con características particulares de diseño y representación.

Para la propuesta de integración de este proyecto se considera exclusivamente el modelo relacional, como requerimiento tanto para los esquemas locales como para los esquemas globales generados. La justificación de la utilización de este modelo, es que sigue siendo uno de los estándares de modelado más utilizado en la industria de las Bases de Datos.

El modelo relacional, basado en la teoría de conjuntos usa como primitiva básica de construcción la relación. Una relación es una tabla

bidimensional. Cada hilera de la tabla contiene datos que pertenecen a alguna cosa o porción de una cosa. Cada columna de la tabla contiene datos sobre atributos. Las hileras son también llamadas tuplas y las columnas atributos

Para que una tabla sea una relación, esta debe cumplir ciertas restricciones. Primero, los atributos deben ser atómicos. Todos los valores en algún atributo deben ser de la misma clase. Cada columna debe tener un nombre único y el orden de las columnas en la tabla es insignificante. Finalmente, dos tuplas en la tabla no pueden se idénticas.

Es importante tomar en cuenta estas consideraciones para la selección apropiada de las BDC´s y para garantizar una consistencia con la metodología de integración propuesta.

3.4 Metodología de integración de esquemas

La necesidad de integrar diversas BDC´s trae consigo la necesidad de utilización de metodologías formales que permitan un proceso de integración confiable y seguro.

Una metodología descompone la integración de esquemas en un número de tareas que pueden ligarse en un proceso interactivo que ofrezca como resultado un esquema global para DBMS federados fuertemente acoplado. Con la idea de mantener un balance entre simplicidad y rentabilidad se describe la metodología de integración propuesta por [Batini y Lenzerini 1986], la cual consiste de cuatro fases: preintegración, comparación de esquemas, adecuación de esquemas y unión de esquemas.

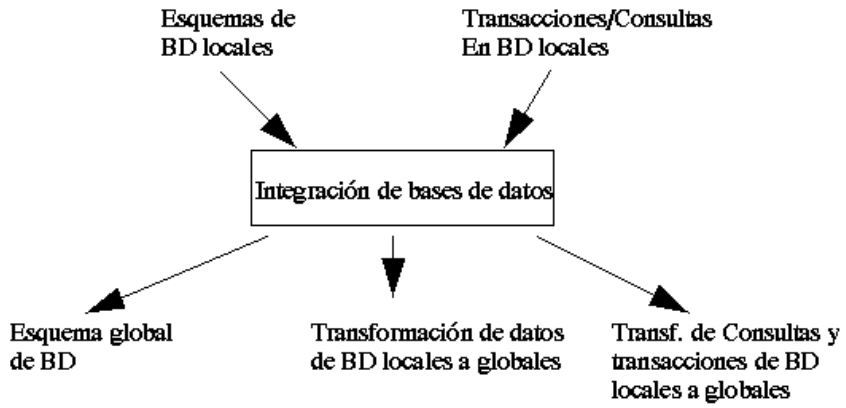

Figura 3.1 Entradas y salidas en la integración de esquemas

3.5.2 Compartición de información

3.5 Preintegración

Otro aspecto importante de esta fase es definir qué información se compartirá y con qué restricciones de acceso. Este proceso podría entenderse como un análogo a la definición de vistas en el modelo relacional. En el contexto de Bases de Datos Federadas, esto puede verse como la determinación de las entidades que cada Base de Datos Componente compartirá con la Federación y su definición en un esquema de componentes.

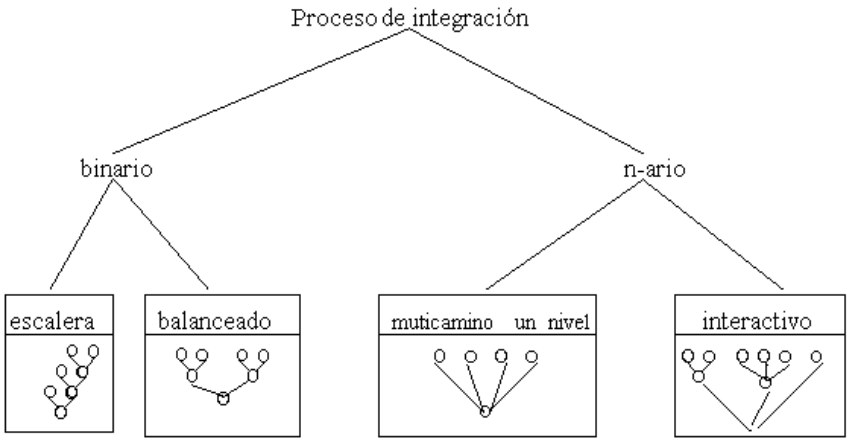

Figura 3.2 Estrategias para el proceso de integración

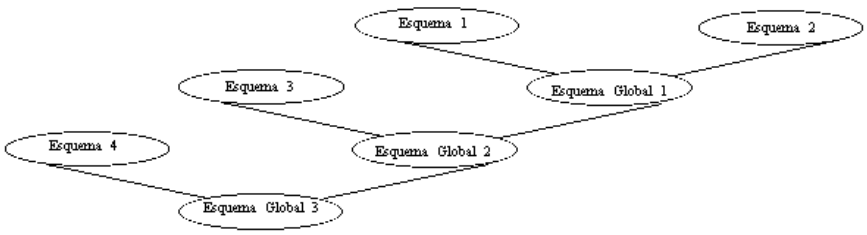

Figura 3.3 Estrategia binaria de escalera para el proceso de integración.

3.6 Comparación de esquemas

En este ejemplo existe la descripción del mismo número de objetos en cada base de datos local, pero a diferente nivel de detalle por lo que es posible establecer la relación lógica a través de un atributo común y ofrecer un acceso global a la información sin redundancias. Ambos tipos de integración consideran información espacial (modelo basado en fragmentos) e información descriptiva.

Esquema Local 1		Esquema Local 2		Esquema Global		
Clave	Objeto	Clave	Objeto	Clave	Objeto	(Esquema Global)
0100	1	1011	2	0100	1	
0101	1	1101	2	0101	1	
:	:	:	:	:	:	
1100	1	1111	2	1100	1	Universo de Información
				1011	2	
				1101	2	
				:	:	
				1111	2	

Esquema Local 1		Esquema Local 2			Esquema Global		
Objeto	Nombre	Objeto	Capital	Sup.	Objeto	Nombre	Sup (Esquema Global)
1	Tlaxcala	1	Tlax.	w m^3	1	Tlaxcala	w m^3
2	Puebla	2	Pue.	x m^3	2	Puebla	x m^3
3	Veracruz	3	Ver.	y m^3	3	Veracruz	y m^3 UNIVERSO DE INF.
:		:	:		:		
n	Coahuila	n	Salt.	z m^3	n	Coahuila	z m^3

3.7 Adecuación de esquemas

La operación inversa, partiendo de la cadena UBICACION [1], puede ser, en algunos casos más difícil. Esto es debido a que en ocasiones es casi imposible proponer los criterios para decidir qué parte de la cadena original corresponde a cada uno de los atributos propuestos para la otra representación. Esto, sin embargo, puede llevar a conflictos con el orden que se usa para cada representación. Debe considerarse , que lo anterior es posible si el sistema soporta solamente consultas, ya que para las operaciones de modificaciones y alta de información, dada la ubicación como un solo atributo, plantea la necesidad de decidir que parte pertenece a cada atributo de la segunda representación.

3.7.4 Claves primarias

El uso de llaves primarias en el modelo relacional permite, en la integración heterogénea establecer la operación de join para integrar los componentes locales. Entonces, será de suma importancia definir cuales son los atributos considerados como claves primarias de cada BDC.

3.7.5 Nivel de precisión

El modelado de datos espaciales a través de la técnica de Quadtree (Ver Apéndice A para detalle de esta técnica) permite representar objetos espaciales a diferentes niveles de precisión. A mayor nivel de precisión, mayor el detalle de representación obtenido del objeto.

Las BDC´s que representan información espacial, pueden manejar niveles

de precisión diferentes, lo que implicaría una manipulación a nivel de implementación para homogeneizar la representación de los objetos espaciales a un mismo nivel de resolución.

Al considerar el nivel de precisión como un conflicto es posible lograr dicha homogeneización a nivel de integración de esquemas.

$$LONGITUD^1_{metros} = \begin{cases} LONGITUD^1_{metros} \ // \text{Función identidad} \\ LONGITUD^2_{millas} * 1.852 \end{cases}$$

$$UBICACION^1 = \begin{cases} UBICACION^1 \\ CALLE^1 \ o \ NUMERO^2 \ o \ COLONIA^3 \end{cases}$$

donde *o* representa la operación de concatenación.

3.8 Unión y reestructuración

Una vez solucionados los conflictos en la fase anterior, es posible la integración de los esquemas. Después de la integración sigue una fase de reestructuración interactiva hasta que se llegue al esquema global deseado. La evaluación de esquema final, es a través de los siguientes criterios:

Completitud y Validez. El esquema integrado debe contener todos los conceptos presentes en los esquemas componentes. El esquema integrado debe ser una representación de la unión de los dominios de las aplicaciones asociadas a los esquemas [Batini y Col 1986].

Representación mínima.- Si el mismo concepto se encuentra en más de un concepto componente, este se debe representar sólo una vez en el esquema integrado.

Comprensibilidad.- El esquema global final debe ser fácil de entender tanto por el diseñador como por el usuario final.

La comprensión detallada de los conflictos de esquema y de datos que se

presentan en este capítulo, permitirá plantear una mecanismo de solución que ofrezca un acceso transparente a las Bases de Datos Componentes. Los conceptos de equivalencia y tipo de integración requieren de atención especial para garantizar una selección apropiada de los componentes a integrar.

La metodología de integración que se discute, establece un patrón de referencia para llevar a cabo la integración de manera semiautomática con la propuesta que se plantea en los capítulos siguientes.