Capítulo 3.
Procesamiento de consultas en un sistema multibase de datos

El procesamiento de consultas en un sistema multibase de datos es la pieza mas importante para la operación del sistema. En este capítulo se describe la arquitectura general de un procesador de consultas multibase de datos. Se mencionan los módulos que integran un procesador de este tipo y la función que deben llevar a cabo.

A grandes rasgos tres pasos son necesarios para procesar una consulta global [Evrendilek y Dogac 1995]: Primero una consulta global es descompuesta en subconsultas de manera que los datos necesitados por cada subconsulta estén disponibles desde cada SBDC (sistema de base de datos componente). Después cada subconsulta es trasladada a una consulta o consultas de el SBDC y enviada (s) al SBDC. Tercero, los resultados retornados por las subconsultas son combinados para dar respuesta a la consulta global.

El procesamiento de consultas es uno de los aspectos mas complejos dentro de un sistema multibase de datos. Aunque esto debiera parecerse a un sistema de bases de datos distribuido existen diferencias debido a que los SBDCs de un sistema multibase de datos normalmente son heterogéneos y poseen distintas capacidades de procesamiento. De esta manera el procesamiento y la optimización de consultas resulta mas difícil que en un sistema de base de datos distribuido.

Las capacidades de procesamiento de consultas de los sistemas de base de datos componentes (SBDCs) pueden variar grandemente, las cuales van desde sistemas de bases de datos orientadas a objetos y sistemas de base de datos relaciónales hasta sistemas de archivos.

El optimizador de consultas global debe descomponer una consulta global en consultas componentes para ser procesadas por los SBDCs. Este también debe determinar como y donde ejecutar algún procesamiento de integración que sea necesario. Para llevar a cabo operaciones de optimización el procesador de la consulta debe de conocer las capacidades de cada SBDC para elegir el mejor plan de ejecución[Attaluri et al. 1995].

3.1.1 Analizador Léxico, Sintáctico y Validación

3.1 Arquitectura de un Procesador de Consultas Multibase de Datos

El analizador léxico identifica los componentes del lenguaje (componentes léxicos) en el texto de la consulta. El analizador sintáctico revisa la sintaxis de la consulta para determinar si está formulada de acuerdo con las reglas sintácticas. Además la consulta se debe validar, para lo cual ha de comprobarse que todos los nombres de atributos y de relaciones sean válidos y tengan sentido desde el punto de vista semántico. Para llevar a cabo la validación este módulo requiere de interactuar con el catálogo del SBDF.

3.1.2 Descomponedor de Consultas

La función del descomponedor es separar una consulta global en unidades de consulta. Una unidad de consulta corresponde a operaciones primitivas necesarias para procesar una consulta, tales como la selección, proyección, o reunión con datos disponibles en la misma base de datos componente. La descomposición puede ser llevada a cabo de acuerdo a las siguientes heurísticas:

1. Selecciones y proyecciones en relaciones sencillas forman unidades de consulta por si mismas.

2. Las operaciones de reunión y las que involucran solamente relaciones almacenadas en la misma base de datos componente también forman unidades de consulta.

3. Cuando una relación es la unión de relaciones en diferentes bases de datos componentes, las unidades de consultas son formadas para cada sitio.

4. Para una reunión (u otra operación) que involucra dos bases de datos diferentes, las relaciones en la condición de la reunión son reemplazadas con unidades de consulta resultantes de las unidades de consulta que recuperan la relación (o parte de esta) desde la base de datos original.

El principio básico aquí es descomponer la consulta en el nivel más fino para explorar todos los planes de ejecución posibles. La información necesaria para poder descomponer la consulta es tomada del catálogo del SBDF.

3.1.3 El Generador de Planes

Dado un grafo de unidades de consulta, el generador de planes construye los planes posibles que consisten de las subconsultas y su secuencia de ejecución. Las unidades de consulta descompuestas son agrupadas para formar subconsultas. Este proceso de agrupación es guiado por las funciones de costo y heurísticas [Elmasri y Navathe 1997]. Un *plan de ejecución de consultas* especifica las subconsultas, los SMBDs componentes involucrados y el tiempo de respuesta esperado el cual es provisto por el evaluador de costo [Lu et al. 1992].

Para determinar un plan de ejecución eficiente, el optimizador de consultas global también necesita estimar los costos de procesamiento de una consulta componente en un SBDC y la cantidad de datos de salida. La cantidad de datos de salida producidos por una consulta componente es un factor decisivo para encontrar un plan eficiente para procesamientos de integración. Debido a que los SBDCs son sistemas pre-existentes autónomos, el optimizador de consultas global no es capaz de obtener la información necesaria de estos para hacer estimaciones exactas [Attaluri et al. 1995].

El generador de planes interactua con el evaluador de costo durante el proceso de generación del plan. Cuando un plan de ejecución es generado, este es pasado al evaluador de costo el cual proporciona un costo estimado. El plan que satisface los objetivos de optimización es finalmente seleccionado como el plan de ejecución para la consulta y enviado al *despachador de subconsultas* . En Lu et al. [1992] y Evrendilek y Dogac [1995] se describen algunas técnicas para llevar a cabo la optimización.

3.1.4 El Evaluador de Costo

El evaluador de costo trabaja conjuntamente con el generador de planes. Su función es estimar el costo de un plan de ejecución de consulta basado en un modelo de costo como el que se describe en la siguiente sección y la información almacenada en el catálogo.

3.1.4.1 El modelo del costo

En la optimización de consultas de un SBDF (sistema de base de datos federada) hay dos objetivos principales: la minimización del tiempo de respuesta y el calculo del costo. El costo de ejecución de una consulta global comprende varios aspectos:

generación de un plan de consulta

invocación de los SBDCs (sistemas de bases de datos componentes)

procesamiento de subconsultas

transferencia de resultados intermedios entre participantes y SBDCs

mediaciones de contexto

ensamble de los resultados en la consulta global

A diferencia de la optimización de consultas distribuidas (en sistemas homogéneos), el costo de una subconsulta no puede ser fácilmente determinado puesto que el optimizador de consultas multibase de datos no tiene información del perfil de la base de datos, las rutas de acceso, o los métodos de acceso que son soportados por los SBDCs. Similarmente el tamaño de los resultados intermedios son generalmente desconocidos y aquí los costos de comunicación y mediaciones de contexto son también difíciles de determinar [Lu et al. 1992].

3.1.5. Despachador de Subconsultas

Se encarga de coordinar la ejecución del plan entre los SMBDs componentes. Establece la conectividad con cada base de datos y le envía las subconsultas que le corresponden, también se encarga de recolectar la información resultante de las subconsultas que posteriormente envía al *combinador de resultados* .

3.1.6 Combinador de Resultados

Lleva a cabo la combinación de los resultados de las subconsultas hechas a cada SMBD componente. En este modulo se deben de combinar la información para resolver selecciones, proyecciones, uniones, reuniones, etc. que involucren mas de un SMBD componente y así dar forma al resultado de la consulta global. Algoritmos para llevar a cabo estas operaciones pueden ser encontrados en [Elmasri y Navathe 1997].

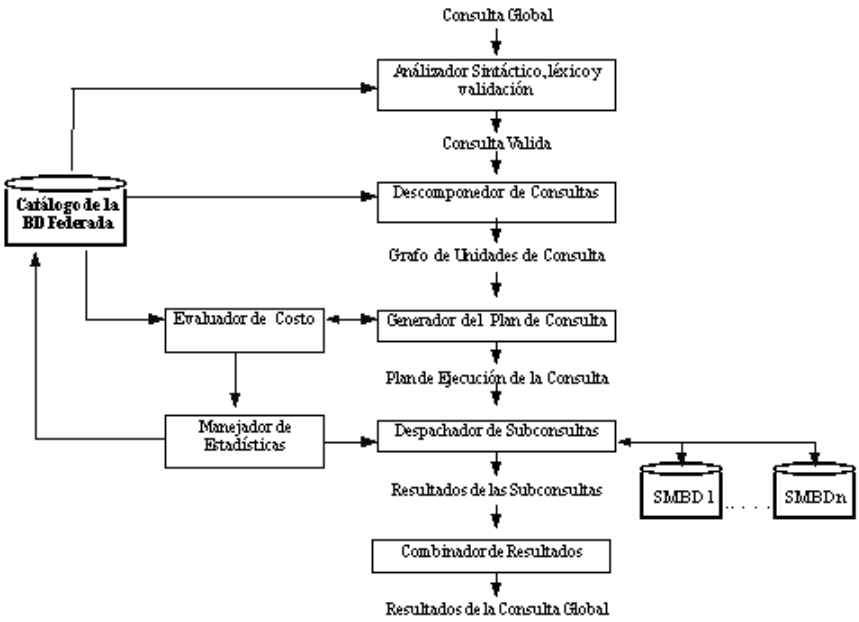

Figura 3.1 Arquitectura de un Procesador de Consultas para un SBDF

3.2 El Catálogo de un SBDF

El catálogo de un SBDF tiene un propósito similar a el catálogo de un sistema manejador de base de datos relacional. Los datos del catálogo son de dos tipos:

1. Datos estructurales. Descripciones de objetos en el sistema y sus relaciones

2. Datos estadísticos. Principalmente estadísticas de los SBDCs utilizados durante la optimización de consultas.

Un SBDC es descrito en el catálogo por propiedades tales como el tipo de la fuente de datos, el modelo de datos usado y las funcionalidades disponibles; los esquemas pueden ser representados en forma relacional, y la información mantenida en el catálogo incluye los nombres de las tablas en el esquema, el nombre y tipo de cada atributo en las tablas, y los mapeos desde el esquema local de los SBDCs a su representación relacional [Attaluri et al. 1995].

Para generar un plan de ejecución para una consulta el generador requiere información de:

La localización de los datos que son referenciados en la consulta

Los nombres de los datos requeridos, así como la manera en que son entendidos por el SBDC.

Estadísticas relacionadas al sistema de los SBDCs, incluyendo

accesibilidad a SBDC (algunos sistemas proveen acceso ilimitado a usuarios foráneos, pero otros proveen acceso limitado a sus recursos).

características de la carga de trabajo medidas en términos de CPU, I/O, y utilización de la línea de comunicación.

Poder de calculo del sistema de computo anfitrión en términos de su velocidad de procesamiento.

Esta información debe ser guardada en el catálogo del SBDF la cual será accesada por el optimizador de consultas.

En una multibase de datos para poder llevar a cabo la optimización de consultas es necesario que el catálogo maneje información estadística referente a los SBDCs. Esta información junto con el tamaño de los resultados intermedios de las subconsultas son los que determinarán la manera en que la subconsulta puede ser optimizada. Sin embargo debido a la heterogeneidad de los SBDCs esta información es difícil de mantener principalmente por las diferentes capacidades de procesamiento de cada SBDC.

En la figura 3.1 pudimos observar que los distintos módulos que integran un procesador de consultas multibase de datos interactuan con el catálogo del SBDF. Por esta razón la funcionalidad optima del procesador de consultas depende en gran medida de la información con la que el catálogo este dotado. Esta información permitirá también determinar si la optimización de consultas puede ser llevada a cabo o no y será pieza fundamental para que los otros módulos que integran el procesador puedan llevar a cabo sus funciones.

En el siguiente capítulo haremos un revisión de los lenguajes para bases de datos y como estos pueden ser implementados.