Ensayos

Ventajas y retos en el uso de bases de datos distribuidas

Resumen

Actualmente diversas actividades en las cuales se encuentran involucradas bases de datos requieren realizarse desde diferentes sitios. Muchas empresas se han diversificado geográficamente y sus recursos de cómputo se encuentran de tal manera esparcidos. Sin embargo, las consultas de datos entre diferentes unidades de negocios son comunes entre ellas, más aún con el advenimiento de la Internet. Las bases de datos distribuidas son una buena alternativa para estos casos. Este documento analiza las principales ventajas de las bases de datos distribuidas y menciona los principales retos tecnológicos en donde aún se está haciendo investigación al respecto.

Abstract

Nowadays, a diversity of activities in which data bases are involved need to be carried out from different sites. Many companies have branched out geographically and likewise their computer resources are spread out. However, consulting common data between different sections of a company is common, and more so now with the arrival of the Internet. Distributed data bases are a good alternative in such cases. This document analyses the principal advantages of distributed data bases and mentions the main technological challenges in which research is still being carried out.

Abstrait

Actuellement des activités diverses dans lesquelles des bases de données sont insérées doivent être effectuées depuis des sites différents. Beaucoup d´ entreprises se sont diversifiées géographiquement et leurs ressources informatiques se sont dispersées. Cependant, les consultations de données communes entre les différentes unités de commerces le sont entre elles, encore plus avec l´ arrivée de l´ internet. Les bases de données distribuées sont une bonne alternative pour tous ces cas. Ce document analyse les principaux avantages des bases de données distribuées et mentionne les défis technologiques principaux où l´ on développe la recherche.

* Francisco de Asís López Fuentes

1. Introducción

El incremento de la globalización y el clima más competitivo ha hecho necesario que las compañías internacionales trabajen de una nueva manera, que maximicen sus sinergias entre sus diferentes unidades de negocios, ingeniería y proyectos alrededor del mundo. Con la explosiva popularidad de la Internet y el world wide web (WWW) hay una necesidad de crecimiento rápido para suministrar acceso sin precedente a fuentes de datos distribuidas globalmente a través de la Internet. La integración de los datos dispersos en diferentes sitios para ser accedidos a través del web, puede requerir de nuevas arquitecturas y herramientas de software para el desarrollo de estos sistemas. Diferentes empresas se han visto en la necesidad de integrarse a estas nuevas tecnologías. Esta necesidad ha creado una fuerte demanda por capacidades de acceso a bases de datos a través de la Internet[1]. En este documento revisamos las ventajas que podemos lograr a través del uso de bases distribuidas, con respecto a una base corporativa centralizada, ambas accedidas a través del web.

2. Arquitectura de una base de datos distribuida

El procesamiento en las bases de datos distribuidas, es el procesamiento por el medio del cual la ejecución de las transacciones, la recuperación y actualización de los datos se lleva a cabo entre dos ó más

* *Profesor Investigador de la Universidad Tecnológica de la Mixteca*

computadoras independientes. La figura 1 muestra un sistema de base de datos distribuida que involucra cuatro computadoras. En esta arquitectura [2] el sistema administrador de base de datos distribuida (DDBMS), esta formado por los administradores de transacciones y los administradores de bases de datos de todas las computadoras.

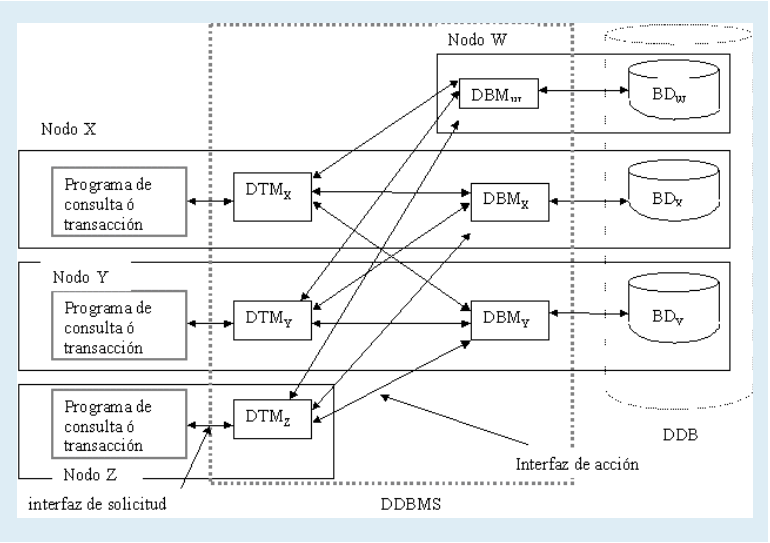

FIGURA 1. ARQUITECTURA DE LAS BASES DE DATOS DISTRIBUIDAS

3. Ventajas para implementar bases de datos distribuidas

La evolución de las bases de datos distribuidas se debe por una parte a razones organizacionales las cuales han demandado que mayores capacidades sean incorporadas a las bases de datos, tales como la integración de información desde distintos sitios donde se encuentre la empresa distribuida hacia algún sitio por ejemplo, para una consulta. Por otra parte, el desarrollo de las tecnologías de comunicación han permitido enlazar datos con aplicaciones que se encuentran en sitios distintos y remotos, por ejemplo las transacciones bancarias realizadas en máquinas-cajeros automáticos (ATM) que se encuentran ubicados en centros comerciales, empresas y escuelas, no serían posibles si no tuviéramos sistemas de comunicación para enlazarnos a bases de datos localizadas en diferentes sitios financieros. A continuación explicamos brevemente en que consisten estas razones:

Económicas y de organización

Muchas organizaciones son descentralizadas y los usuarios de los sistemas de información en estas corporaciones como en los bancos, grupos industriales, servicios nacionales de salud y educación ven más útil un enfoque de base distribuida que refleje la estructura de la organización [7]. Esto ha podido ocurrir con el desarrollo reciente de tecnologías de cómputo, la presión ejercida por los usuarios y el advenimiento de las nuevas tecnologías de comunicación.

Interconexión entre bases de datos existente

Surge ante el planteamiento de un modelo poco óptimo para aquellas empresas en donde las fuentes locales de información son las principales usuarias de su propia información, es decir un departamento necesita un software y hardware específico que pocas veces cruza datos con otros departamentos. Este enfoque aislado trae mejores beneficios de seguridad y disponibilidad de la información, ya que la implantación de los mecanismos de control de acceso fue más fácil. Sin embargo cuando existió la necesidad de transferir datos entre los diferentes sistemas departamentales y el sistema central de una corporación hubo problemas de consistencia y duplicidad. Las bases de datos distribuidas son la solución natural cuando diversas bases de datos existen en una organización y se tiene la necesidad de ejecutar aplicaciones globales. En este caso, la base de datos distribuida es creada por encima de las bases de datos locales preexistentes, lo cual puede requerir un cierto grado de reestructuración local [7]. Esto permite un tipo de control centralizado o distribuido para mantener la integridad de la base de datos descentralizada en diferentes sitios. La descentralización aunque, en un principio concebida para sistemas homogéneos de información, involucra en varios casos el manejo de sistemas heterogéneos. La heterogeneidad se puede dar a muchos niveles, desde la forma de concebir los datos hasta los medios de almacenamiento para mantener su durabilidad, pasando por los diferentes sistemas de comunicación para transportar los datos.

Crecimiento proporcional

Existen organizaciones que crecen adicionando nuevas unidades organizacionales relativamente autónomas como: sucursales, nuevos almacenes o fábri-

cas, lo que implica el desarrollo de nuevas bases de datos para los sistemas de información. Es aquí donde un enfoque de bases de datos distribuido es muy útil, ya que estas soportan un suave crecimiento con un mínimo impacto en las unidades existentes alrededor. En un enfoque centralizado, aún en las dimensiones del sistema inicial se tiene cuidado con futuras expansiones, se dificulta y se encarece al implementarlo y el crecimiento por nuevas aplicaciones afecta también a las aplicaciones ya existentes.

Reducción de la sobrecarga de comunicación

Cuando existen aplicaciones distribuidas geográficamente en distintos nodos de una red, un enfoque bajo una base de datos centralizada lleva a aumentar el tráfico en la red, dificultando los accesos a la base de datos. Un enfoque distribuido permite reducir la sobrecarga de tráfico en la red ya que los sitios locales pueden contener los fragmentos más usados por las aplicaciones locales, esta ventaja se debe considerar al diseñar la base de datos distribuida.

Consideraciones de rendimiento

La existencia de diversos procesadores autónomos resulta en el incremento del rendimiento a través de un alto grado de paralelismo. Esta consideración puede ser aplicada a cualquier sistema de multiprocesador y no solamente a bases de datos distribuidas. Sin embargo, las bases de datos distribuidas tienen la ventaja de que la descomposición de datos refleja el criterio de dependencia de aplicaciones lo cual maximiza la situación de las aplicaciones; de esta manera la interferencia mutua entre diferentes procesos es minimizada. La carga es compartida entre los diferentes procesadores y los cuellos de botellas críticos tales como: la misma red de comunicación o servicios comunes del propio sistema se evitan. Este efecto es una consecuencia de la capacidad de procesamiento autónomo requeridos, para las aplicaciones locales, declarada en la definición de las bases de datos distribuidas.

Confiabilidad y disponibilidad

El enfoque de bases de datos distribuidas, especialmente cuando se tiene redundancia de datos, puede ser usado también con el fin de obtener una mayor confiabilidad y disponibilidad. Sin embargo obtener esta meta no es sencillo y requiere el uso de técnicas complejas. Las fallas en un sistema distribuido se pueden dar con mayor frecuencia que en un sistema centralizado, debido al gran número de componentes que lo integran, sin embargo el efecto de cada falla se reduce a aquellas aplicaciones que usan el dato y el sitio que falla, y el hecho- de que una falla en un sitio o en un dato ocasione que el sistema distribuido completo falle, es muy raro.

Últimamente los sistemas distribuidos están siendo afectados aún más por el desarrollo dramático de los canales de distribución de la información basadas en la penetración de la Internet y a su rápido avance en áreas no asociadas normalmente a la computación [3]. El World Wide Web, el E-mail, y los grupos de Internet son ejemplos prominentes. Esta tendencia no sólo continuará, sino que se acelerará.

Actualmente existen aplicaciones que desde su origen han sido concebidas como distribuidas, donde se han considerado las diferentes tecnologías recientes que permiten integrar los sistemas de información sin afectar al usuario. Sin embargo, un reto importante es cuando tenemos que integrar a un ambiente distribuido diferentes departamentos, donde existen sistemas gestores de bases de datos (DBMS) locales distintas (heterogéneas), que gestionan datos que se requieren cruzar en diferentes áreas para aplicaciones distintas.

Con el comercio electrónico que comienza a ser una característica común de algunas empresas, la importancia de los sistemas distribuidos y las bases de datos distribuidas se acrecienta, ya que aunque actualmente la mayoría de los diseñadores e implementadores de los sistemas de comercio electrónico se han concentrado en lo que respecta al pago electrónico, en realidad existen muchos otros componentes en la implementación de sistemas distribuidos masivos. El comercio electrónico implica no solamente el pago por las mercancías o servicios [6] sino también su creación, publicidad, salida, mantenimiento y disposición.

4. Los retos tecnológicos

Como expectativa con respecto a las tecnologías de DBMS distribuidas y paralelas, hay un número de cuestiones que tienen todavía que ser resueltas satisfactoriamente. Algunas de estas cuestiones de investigación importantes, [4] son indicadas a continuación:

Colocación del dato

En un sistema paralelo, la colocación apropiada de los datos es esencial para balancear la carga. Idealmente, la interferencia entre las operaciones paralelas simultáneas se puede evitar teniendo cada trabajo de la operación sobre un conjunto de datos independiente. Estos conjuntos de datos independientes pueden ser obtenidos por desagrupar (dividiendo horizontalmente) las relaciones según una función (función hash o índice del rango) aplicada a algún atributo(s) de la colocación, y asignando cada partición a un disco diferente. Como con la fragmentación horizontal en bases de datos distribuidas, el desagrupamiento es útil para obtener paralelismo entre consultas, teniendo consultas independientes trabajando en particiones diferentes, y paralelismo entre consultas, por tener una operación de consulta trabajando en diferentes particiones. El desagrupamiento puede ser de un sólo atributo o de muchos atributos. En el último caso, una consulta igual requerirá la igualación de todos los atributos que se puedan procesar por un solo nodo sin comunicaciones. La selección entre el hashing y el índice del rango para repartir es una cuestión del diseño: el hashing incurre en menos gastos de almacenaje pero proporciona únicamente ayuda directa para las consultas igualmente exactas, mientras que el índice del rango puede también utilizar consultas del rango. Propuesto inicialmente para sistemas no compartidos, el desagrupamiento ha mostrado también ser útil para el diseño de memoria compartida, por reducir conflictos de acceso a memoria. El desagrupamiento completo, por lo cual cada relación se reparte a través de todos los nodos, causa problemas para la relación o los sistemas pequeños con una gran cantidad de nodos [4]. Una solución mejor es el desagrupamiento variable, donde cada relación se salva en cierto número de nodos como una función de la frecuencia de acceso y del tamaño de la relación. Esto puede ser combinado con el agrupamiento de múltiples relaciones para evitar la carga general de comunicación de las operaciones binarias. Cuando los criterios usados para la colocación de los datos cambian hasta el punto de que el balanceo de la carga se degrade perceptiblemente, la reorganización dinámica es requerida. Un problema serio en la colocación de los datos es cómo tratar con las distribuciones sesgadas de los datos lo cual puede conducir a una repartición no uniforme y afectar negativamente el balanceo de la carga. Un factor de complicación final en la colocación de los datos es la replicación de los datos para una alta disponibilidad. Un enfoque ingenuo [2] es mantener dos copias de los mismos datos, una copia primaria y una de respaldo, en dos nodos separados. Sin embargo, en caso de una falla del nodo, la carga del nodo que tiene la copia puede duplicarse, de tal modo que afecte el balanceo de la carga.

Problemas de escalamiento y fallas de la red

La comunidad de base de datos no tiene un completo entendimiento de las implicaciones de rendimiento de todas las alternativas de diseño que acompañan el desarrollo de DBMS distribuidos. Específicamente cuando nos referimos a la escalabilidad de algunos protocolos y algoritmos, cuando el sistema llega a ser distribuido geográficamente o cuando el número de componentes del sistema se incrementa. De preocupación específica podemos mencionar los mecanismos de procesamiento de transacción distribuida en sistemas de base de datos distribuidas basados en redes WAN. Diversos algoritmos y protocolos propuestos para arquitecturas de redes de área local (LAN), no están bien entendido su comportamiento cuando son llevados a redes de área amplia (WAN) [6]. Para lo cual se requiere una mayor investigación con respecto a los modelos de rendimiento. También el aislamiento y la corrección de fallas en los sistemas distribuidos requerirán nuevos servicios de la infraestructura para vigilar la calidad de las comunicaciones y entregar avisos de las anomalías a los proveedores del servicio cuando la calidad se sitúa por debajo de un umbral dado [1]. Posibles soluciones serían la implementación de mecanismos automáticos de aislamiento y detección de fallas.

Procesamiento de consultas paralelas y distribuidas

La optimización de la consulta global genera un plan óptimo de ejecución para la consulta del fragmento de la entrada de información tomando decisiones con respecto al orden de la operación, al mover el dato entre los sitios, y a la selección tanto de los algoritmos distribuidos y de los locales, para las operaciones de la base de datos. Hay varios problemas relacionados a este paso. Se tienen que hacer con las restricciones

impuestas ante el modelo de costo, la concentración en un subconjunto del lenguaje de consulta, la negociación entre el costo de la optimización y el costo de la ejecución, y el intervalo optimización-reoptimización. El modelo de costo es central para optimización de consultas globales, ya que proporcionan la abstracción necesaria del sistema de ejecución del DBMS distribuido en términos de acceso, tanto como en la abstracción de la base de datos en términos de información del esquema físico relacionada estadísticamente. El modelo de costo es usado para predecir el costo de ejecución de los planes de ejecución alternativos para una consulta. Un número importante de restricciones son frecuentemente asociadas con el modelo de costo [4], los cual limitan su eficiencia de optimización para mejorar el rendimiento efectivo. Es necesario negociar entre el costo de optimización y la calidad del plan de ejecución generado. La optimización de la consulta global se realiza típicamente antes de la ejecución de la consulta; de aquí que esta sea llamada estática. Un problema importante con este enfoque es que el modelo de costo usado para la optimización puede llegar a ser inexacto, debido a cambios en el tamaño del fragmento o a la reorganización de la base de datos que es importante para el balance de la carga.

Procesamiento de transacciones distribuidas

Existen aún tópicos de fuerte investigación en el área de procesamiento de transacciones distribuidas. Con respecto a la replicación de datos, la investigación requerida se encamina a los métodos de replicación para computación y comunicación; y más trabajo es requerido para permitir la explotación sistemática de las características de aplicaciones específicas. Una de las dificultades en las técnicas de replicación de evaluación cuantitativa yace en la ausencia de modelos de incidencia de falla comúnmente validados. Los modelos de Markov que son algunas veces usados para analizar la disponibilidad alcanzada por los protocolos de replicación asumen la independencia estadística de los eventos individuales de falla y la muy rara división de la red por causas relacionadas a fallas en los sitios. Sin embargo, actualmente no se conoce que una u otra de estas suposiciones sea alcanzable, tampoco se conoce como responden los modelos de Markov a estas suposiciones. Los modelos de Markov [6] para su simulación requieren mediciones empíricas, debido a que las simulaciones frecuentemente incorporan las mismas suposiciones que sirven de base al análisis de Markov. Hay una necesidad, por lo tanto, de estudios empíricos para monitorear patrones de fallas en sistemas de producción de la vida real, con el propósito de construir un modelo simple de carga de fallas típicas.

Heterogeneidad

Esta cuestión importante en el diseño actual de las bases de datos distribuidas será perceptiblemente mayor para sistemas distribuidos masivos. Mientras que la mayoría de las aplicaciones distribuidas existentes [5] se ejecuta en una cantidad de diferentes plataformas de cómputo, limitadas a un pequeño número de familias comunes, por ejemplo UNIX, Windows, LINUX o quizás MVS. Las aplicaciones distribuidas masivas, por otra parte, se ejecutarán no solamente en plataformas existentes [1], sino también en una amplia variedad de sistemas empotrados, soportados por los propios sistemas operativos y hardware del propietario (tal como sistemas de control en automóvil y PDAs). Una aplicación distribuida masiva para la comunicación remota podrá tener componentes que se ejecuten en estaciones de trabajo, en equipos para TV por cable, en teléfonos portátiles, en dispositivos de comunicación basados en PCS y así sucesivamente [6]. Esto aumentará el número de diversas implementaciones de software para un solo tipo de componente, de modo que será necesario un esfuerzo para asegurarse de que la aplicación trabaja correctamente en un ambiente heterogéneo.

Representación, codificación y traducción de objetos

Hay una variedad de esfuerzo para determinar los mejores modelos de programación para los objetos distribuidos, tales como CORBA y Java. Sin embargo, existen ciertas cuestiones que introducirán nuevos retos en cómo se representan, se codifican y se traducen los objetos. La representación de objetos distribuidos masivos requerirá no solo nuevas técnicas, sino que su presentación a los usuarios también requerirá innovación. Algunos investigadores han examinado este problema. Una nueva clase de interfaz de usuario representa objetos como espacios virtuales [2]. Esta técnica es conveniente para presentar objetos distribuidos masivos a los usuarios finales. Por ejemplo,

un objeto de primer nivel se puede representar como un mundo virtual, sus componentes se ponen como países, ciudades, calles, casas, recámaras, etc., la estructura exacta dependerá del tamaño del objeto de primer nivel y su interrelación con los componentes así como la interrelación entre los mismos componentes. Tales paradigmas de la presentación serán requeridos para que los objetos distribuidos masivos sean accesibles al usuario.

Administración de recursos

Con el fin de diseñar y construir aplicaciones masivas distribuidas, los ingenieros tendrán que enfrentarse con nuevos problemas en la administración de recurso. Muchos sistemas distribuidos existentes funcionan según un modelo local de control de recursos. El proceso local maneja sus propios recursos, obrando recíprocamente con otros hilos de control a través de métodos de invocación de paso de mensajes o RPC [3]. En sistemas distribuidos masivos, los objetos estarán compuestos de recursos situados en una gran cantidad de distintos lugares. Controlar los recursos asociados a un objeto, solamente será posible a través de un mecanismo global distribuido de administración de recurso del objeto. Esto introducirá nuevas cuestiones en el control de los recursos del sistema distribuido:

Protección

La protección de los recursos distribuidos del sistema, que incluye recursos básicos tales como procesadores, almacenamiento, comunicaciones, E/S así como los componentes de alto nivel de estos recursos (tales como procesos, archivos, mensajes, ventanas de visualización y objetos más complejos) no es un aspecto que se tenga solucionado aún en los sistemas distribuidos existentes [6]. Mientras los ingenieros están actualmente ocupados en desarrollar soluciones para los muchos problemas que existen en esta área, no están tratando las cuestiones de la protección que se presentarán, si los sistemas distribuidos llegan a ser muy grandes. Los sistemas distribuidos masivos en su mayor parte soportarán una gran cantidad de sistemas terminales, muchos de los cuales serán empotrados en otros equipos y usados por clientes tecnológicamente ingenuos. La escala de los sistemas distribuidos introducirá nuevos problemas en el control de acceso al recurso para los sistemas terminales y los sistemas de ayuda de la infraestructura. Los implementadores requerirán técnicas nuevas, tales como jerarquías de la lista de acceso, sistemas que solucionen la revocación y los problemas de objetos y técnicas de control de acceso que combinen las ventajas y características de las listas de control de acceso, con los controles de acceso por capacidades y eliminen las desventajas de cada uno. Cuando nos enfrentamos con el problema de la protección de la información, los sistemas distribuidos existentes deben hacer frente a los controles que los gobiernos han puesto en tecnología criptográfica [3]. Esto ha obstaculizado a los ingenieros en proporcionar niveles de seguridad apropiados a los usuarios de sistemas distribuidos.

Conclusiones

Mostrar las ventajas que tienen la implementación de bases de datos distribuidas resulta importante para tener conocimiento de cómo la información para diferentes aplicaciones en ingeniería [5] y de negocios se puede distribuir y replicar en diferentes sitios cuando ciertos sitios locales tienen capacidades de almacenamiento y procesamiento limitadas, pero tienen la ventaja de poder integrarse a otros sitios remotos con mejores recursos por medio de una red de comunicación. Las redes de comunicación son un punto fundamental para que las bases de datos pasen de un escenario centralizado a uno distribuido. El uso de bases de datos distribuidas nos permite poder escalar nuestros recursos de cómputo en forma paulatina sin tener que necesariamente adquirir un sistema nuevo completo. Sin embargo, aún existen áreas en las bases de datos distribuidas que se encuentran en investigación y desarrollo, las cuales son un reto tecnológico para varios grupos de investigadores. En este documento mencionamos algunas de ellas, tal como es la localización del dato, la replicación de fragmentos, la tolerancia a fallas en la red o la seguridad, esto con el fin de dar al lector interesado en las bases distribuidas, un punto de referencia de los temas actuales con respecto a éstas

Bibliografía

[1] Tam Nguyen and V. Srinivasan.

1996 Accesing relational database from the world wide web. In proceeding of the 1996 ACM-SIGMOD Conference, pages 529-540, Montreal,Canada.

[2] Baker M. Scott and Moon Bongki.

1999 Distributed Cooperative Web Servers, 8th International World Wide Conference, Toronto Canada.

[3] Douglas Comer

1997 Redes de computadoras, Internet e Interredes, 1ª. Edición Prentice-Hall.

[4] Öszu Tamer y Valduriez Patrick.

1999 Principles of Distributed Database Systems, Prentice-Hall.

[5] López-Fuentes F., Ramos P. Erik.

2001 Cómputo distribuido para el estudio de flujos de carga en los sistemas eléctricos de potencia, 1º. Congreso de Ingeniería Electrónica y Computación.

[6] Chen J; DeWitt; Niagara C. Q. A .

Scalable Continuos Query System for Internet databases

[7] Ceri y Pelagatti.

1985 Distributed database, McGraw.Hill.

Nota Aclaratoria

En el ensayo "Efectos de la fertilización nitrogenada y la biofertilización en la calidad y conservación postcosecha del tomate" publicado en el número 17, se dice que la coautora María Isabel Hernández Díaz es profesora de la Universidad Tecnológica de la Mixteca, cuando en realidad todos los autores de este artículo, laboran en el Instituto de Investigaciones Hortícolas "Liliana Dimitrova" en La Habana Cuba.