Investigación de Bases de Datos: Logros y Oportunidades en el Siglo 21

Gonzalo Mena Mendoza
gonzalo@mena.com.mx
Maestría en Ingeniería de Software Distribuido
Facultad de Informática, Universidad Autónoma de Querétaro
17 de septiembre de 2005

Artículo Original

Silberschatz, Stonebraker, Ullman (editores). "Database Research: Achievements and Opportunities Into the 21st Century." Report on a NSF Workshop on the future of Database Systems Research, mayo 26-27, 1995.

Introducción

Puntos del taller de 1995 de la NSF sobre investigación en base de datos:

- La comunidad de investigadores de bases de datos juegan un papel fundamental en crear la infraestructura tecnológica sobre la que los avances de bases de datos evolucionan.
- Explosión de información digital. Grupos de nuevos problemas: objetos multimedia, distribución de información, nuevas aplicaciones, gestión de flujos de trabajo y de transacciones, facilidad de gestión y uso de la base de datos.
- Explosión en la capacidad de almacenamiento, procesamiento y comunicación.
- Necesidad de apoyo gubernamental e industrial para la investigación básica de bases de datos.

Nuevos aplicaciones demandan nuevas soluciones. La comunidad de investigadores tiene un buen historial en crear nuevas tecnologías y ponerlas en práctica. Es una buena inversión mantener una comunidad de investigación saludable y viable.

Se observa un patrón consistente entre la construcción de teorías, la derivación de principios funcionales que conducen a estudios experimentales y la contrucción de prototipos, que a su vez se convierten en productos comerciales.

Logros recientes

- **Bases de datos orientadas a objetos** (OODB) y de objetos relacionales (ORDB).
- **Nuevos tipos de datos**: espaciales y temporales.
- **Procesamiento de transacciones**: replicación, transacciones "largas", manejo de "versiones y configuraciones".

Nuevas aplicaciones

- **EOSDIS** (Earth Observing System Data Information System): 1 petabyte recolectado cada 3 años.
- **Comercio electrónico**: muchos participantes que desconfían de los otros, fuentes de datos, heterogeneas, autentificación y transferencia de fondos distribuidas.
- **Sistemas de información de salubridad**: fuentes de datos heterogeneas, confidencialidad de la información, interfaces apropiadas para personal de salubridad.

- **Publicación digital**: gestión y entrega de enormes cuerpos de datos a gran velocidad, posiblemente con restricciones de tiempo real, protección de propiedad intelectual, organización y acceso a volúmenes enormes de información.
- **Diseño colaborativo**: integración de fuentes heterogeneas, nuevas formas de concurrencia y mecanismos para compartir, gestión de flujos de trabajo en los que transacciones "largas" interactuan de manera sólida, soporte de versiones del mismo componente y manejo de configuraciones que combinan versiones de muchos componentes.

Tendencias que afectan la investigación de bases de datos

- **Tendencias tecnológicas**: aumentos en un factor de 10 cada diez años en MIPS, costo de procesador típico, cantidad de almacenamiento secundario, cantidad de memoria principal: Otros factores que no han crecido tan rapidamente: bits transmitidos por unidad de costo, bits transmitidos por segundo. Esto genera un ambiente en el que es posible realizar consultas complejas sobre terabytes de datos de manera barata.
- **Tendencias de arquitectura de bases de datos**: bases de datos relacionales ubicuas, aplicaciones cliente-servidor basadas en bases de datos relacionales, modelos de datos más ricos que los méramente basados en registros.
- **Clima de investigación y de negocios**: las grandes corporaciones y agencias gubernamentales que sostenían investigación básica o a largo plazo han hecho reducciones hacia investigaciones a corto plazo. Sin embargo el crecimiento de la industria de la información ha puesto a las bases de datos en un lugar prominente de las preocupaciones corporativas.
- **Internet**: explosión demográfica y de información disponible. Los gestores de bases de datos ocupan un lugar importante en muchos de los principales sitios de la web.

Nuevas direcciones de investigación

Manejo de objetos multimedia

- **Almacenamiento terciario**: debido al enorme tamaño de los datos multimedia, estos no se almacenarán en discos magnéticos, sino en dispositivos de almacenamiento "terciario" como torres de Cds y silos de cintas que son órdenes de magnitud más lentos que los discos magnéticos.
- **Nuevos tipos de datos**: cada forma de información multimedia requiere de sus propios operadores y funciones. Además será necesaria la integración de datos que involucren distintos de estos tipos.
- **Calidad de servicio**: cómo asegurar la presentación realista y oportuna de los datos, cómo degradar con "gracia", cómo interpolar o extrapolar los datos.
- **Consultas con múltiples respuestas**: nuevos lenguajes de consulta o extensiones para especificar grados de precisión, por ejemplo en la búsqueda de imágenes por la que mejor "pega" o por características imprecisas como forma, color o textura.
- **Manejo de interfaz del usuario**: SQL no es una buena herramienta para consultas geográficas. Nuevas formas de realizar consultas, distintas maneras de consultar el mismo medio.

Distribución de información

- **Grado de autonomía**: los sistemas distribuidos son muchas veces propiedad de distintos participantes. Cómo manejar participantes que niegan una conexión o que tienen sistemas con distintas capacidades.
- **Contabilidad y facturación**: algoritmos y mecanismos para cobrar por información distribuida entre distintos participantes. Micropagos.
- **Seguridad y privacidad**: sistemas de autentificación y autorización extremadamente flexibles.

Mecanismos que permitan la venta de información a grandes cantidades de usuarios cuya identidad desconoce el vendedor.

- **Replicación y reconciliación**: los sistemas deben operar tan bien como puedan cuando sus componentes pierden la conexión temporalmente. Desarrollar algoritmos para reconciliación de datos.
- **Integración y conversión**: cuál debe ser el modelo de integración. Herramientas para utilizar de fuentes de datos disímbolas tan fácilmente como si fueran bases de datos unificadas. El "problema de ontología" de la inteligencia artificial.Utilización de mediadores
- **Recuperación y descubrimiento**: la naturaleza de la web, una colección informal de recursos enlazados, presenta problemas como el manejo de datos cuya esquema no es claro, cambia sin aviso y tiene estructura irregular, datos con precisión o confiabilidad no clara, manejo de información semiestructurada.
- **Calidad de datos**: métodos para evaluar la calidad de la información proveniente de distintas fuentes. Capacidad de consultar la confiabilidad del "linaje" de los datos.

Nuevos usos

- **Minería de datos**: extracción de información de cuerpos enormes de datos. Las consultas tienden a ser ad hoc. Muchas veces la pregunta es escontrar "algo interesante". Necesidad de optimización de consultas complejas sobre datos multidimensionales, técnicas de optimización de almacenamiento terciario, interfaces de consulta de alto nivel.
- **Almacenes de datos**: almacenamiento físico de información integrada. Necesidad de herramientas que "bombeen" información de datos operativos hacia el almacén, métodos para "tallar" información, facilidades para mantener un metadiccionario con el origen de los datos.
- **Repositorios**: gestión y almacenamiento de datos y metadatos, por ejemplo herramientas CASE. Manejo de versiones y configuraciones. La meta es crear "sistemas de gestión de repositorios".

Gestión de transacciones y flujos de trabajo

- **Gestión de flujos de trabajo**: los procesos de negocio involucran pasos realizados por la computadora y otros realizados por humanos, es necesario tipos especiales de manejo de datos que tomen en cuenta secuencias de eventos relacionados que pueden involucrar disyuntivas y desandado.
- **Modelos alternativos de transacciones**: las transacciones requieren de atomicidad, serializabilidad y recuperabilidad. Se han propuesto otros modelos que involucran transacciones anidadas.

Facilidad de uso

Debido al papel de la información en la sociedad se expande rápidamente, se requiere de mejores interfaces, no sólo para el usuario final, sino para simplificar las tareas de instalación, actualización y afinación.

Conclusiones

Debido a que las demandas actuales de información están probando los límites de la tecnología de bases de datos, es necesario que la comunidad de investigadores se afronte rápidamente estas oportunidades que van desde fundamentos teóricos de nuevos modelos y algoritmos hasta la creación de ambiciosos prototipos. Sin embargo el financiamento de esta investigación está muy por abajo de aquel de otras areas de importancia comparable.

Por lo que se reitera la necesidad de promover el involucramiento de las agencias gubernamentales encargadas de evaluar y promover la investigación de bases de datos así como de las empresas comerciales que son beneficiadas por dicha investigación. Anticipan una década fructífera en logros

tanto académicos como industriales y esperan la respuesta proactiva de gobierno y empresas.

Comentarios

El contenido general del reporte sigue siendo vigente a diez años de su publicación. Algunos logros mencionados en aquel entonces, como las bases de datos orientadas a objetos no han tenido el auge que se pensó y las bases de datos relacionales siguen reinando con la incorporación de mejores funcionalidades de replicación, distribución, disponibilidad, etc. Incluso bases de datos de nuevo cuño y libres como MySQL y Postgres cuentan con propiedades transaccionales, de replicación y manejan tipos de datos geospaciales.

Las tendencias tecnológicas permanecen avanzando al mismo ritmo, quizás con un sesgo hacia los sistemas multiprocesador debido a las limitantes de velocidad de los procesadores convencionales. El clima de recortes en la investigación básica y a largo plazo es muy similar, basta mencionar la clausura de los famosos laboratorios Bell hace unos días.

Algunas predicciones como la del empleo difundido de almacenamiento terciario no se han cumplido de manera puntual, eso último debido al gran abaratamiento de los medios de almacenamiento convencionales. Por ejemplo, las consultas en Google que involucran volúmenes de información del órden de petabytes se realizan sobre RAM, ni siquiera sobre discos magnéticos. Otras como los mecanimos de micropagos no han tenido éxito para proveedores individuales y queda por saber si serán prácticos para agregadores de información de distintas fuentes.

Sin embargo el uso de herramientas de minería de datos y data-warehousing han encontrado gran aceptación, así como las bases de datos geoespaciales. Se han encontrado mecanismos más o menos buenos para la evaluación y consulta de grandes volúmes de información semiestructurada aunque la “web semántica” continúa siendo fundamentalmente una buena idea por realizarse.

Muchas de las oportunidades de investigación continúan llenas de problemas abiertos interesantes, como los de seguridad, privacidad, reconciliación, integración y consulta difusa. Y otras tendencias no vislumbradas en 1995 han surgido, como la representación e intercambio de información mediante XML.