Teórico 1 – Introducción a las bases de datos

¿De qué trata este curso?

Este curso pretende brindar un primer acercamiento a las bases de datos, y en especial a las bases de datos relacionales.

Se pretende que al finalizar el mismo el estudiante sea capaz de comprender los conceptos fundamentales del modelo relacional, dominar técnicas de diseño tanto conceptual como lógico de una base de datos relacional, y realizar consultas sobre la misma.

¿Qué es una base de datos?

Definiciones

Clásicamente, se define por base de datos un conjunto de datos relacionados entre sí. Con semejante definición los límites de nuestro objeto de estudio quedan muy vagamente definidos. Se supone implícitamente que una base de datos cumple además con las siguientes tres propiedades:

- Representa algún aspecto del mundo real
- Es un conjunto de datos lógicamente coherente
- Se diseña, construye y puebla para un propósito específico

El software que maneja los datos de una base de datos se conoce como manejador de base de datos o DBMS (Database Management System), algunas veces llamado motor de base de datos (database engine). Un DBMS facilita el trabajo de definir, construir y manipular bases de datos para diversas aplicaciones.

No todas las aplicaciones que manejan datos se beneficiarían al utilizar un DBMS. En muchos casos donde la información no cambia, la cantidad de datos es pequeña o su estructura muy simple, los datos pueden mantenerse en memoria o se pueden utilizar archivos.

Un DBMS supone una tecnología madura para la creación y manipulación de una base de datos, que permite delegar muchos de los problemas relativos al manejo de los datos (como restricciones de integridad, múltiples vistas, seguridad, concurrencia y posibilidades de recuperación).

Sin embargo, utilizar un DBMS puede implicar costos de inversión en hardware, software y capacitación, y costos de tiempo invertido en administración del DBMS y de las bases de datos.

Usualmente se distingue entre el esquema de la base de datos, que define las estructuras y suele ser bastante estable; de la instancia de la base de datos que consiste de los datos y suele ser muy inestable. Un DBMS ofrece lenguajes para crear las estructuras (DDL, Data Definition Language), manipular los datos (DML, Data Manipulation Language) y para realizar consultas (QL, Query Language).

Algunos conceptos quedarán claros al ver su aparición en una breve cronología de los eventos más importantes relacionados al desarrollo de la teoría de las bases de datos, y a la construcción de DBMS reales.

Un poco de historia

Durante los 60, las computadoras se habían vuelto accesibles para organizaciones de los gobiernos y empresas de gran tamaño como los bancos y las compañías aéreas. Pronto, todos los usuarios de computadoras que almacenaron grandes volúmenes de información, comenzaron a tener los mismos problemas para manejar tantos datos, por lo que había gran interés por la investigación en estos temas.

En esta época se desarrollaron dos modelos para el manejo de grandes bases de datos, un modelo de base de datos en red (CODASYL) y un modelo jerárquico (IMS). El objetivo de ambos era desarrollar un lenguaje común para el manejo de registros.

Se creó una comisión CODASYL (compuesta principalmente por fabricantes de computadoras y el departamento de defensa de Estados Unidos) que participó en la creación del lenguaje COBOL (COmmon Business-Oriented Language), muy utilizado hasta hoy en día en empresas de gran porte, sobre todo en ambiente mainframe.

El modelo jerárquico IMS (Information Management System) es un refinamiento del modelo de base de datos en red. Fue desarrollado principalmente por IBM en el marco del programa Apollo de la NASA, y se destacan sus capacidades de indexación. Los registros en una base jerárquica se almacenan en una estructura de árbol, con relaciones de tipo padre-hijo, y las relaciones entre los registros se establecían a nivel físico (sectores y pistas).

El modelo jerárquico tiene muchas limitaciones, sobre todo en lo que respecta a control sobre los datos como control de registros duplicados o integridad referencial. Sin embargo, se utilizan bases de datos jerárquicas hasta la actualidad, especialmente Adabas (Adaptable Database System) accedidas por aplicaciones desarrolladas en lenguaje NATURAL.

El "prócer" de la disciplina del manejo de bases de datos fue Edgar Frank Codd (1923 - 2003) quien propuso el modelo relacional en 1969, en un reporte de investigación de IBM llamado "Derivability, Redundancy and Consistency of Relations Stored in Large Data Banks". Este trabajo fue crítico con el enfoque existente hasta el momento, que ataba la descripción de la información a los mecanismos de acceso.

El trabajo de Codd, con un enfoque mucho más basado en la lógica y en la matemática, ha sido la piedra filosofal de toda la teoría de bases de datos relacionales. Se denomina RDBMS a un DBMS relacional.

A principios de los 70, Chamberlin y Boyce (también investigadores de IBM) crearon un lenguaje de consultas llamado SEQUEL, que después pasaría a llamarse SQL.

Tras doctorarse en 1971, Michael Stonebraker realizó investigaciones para IBM hacia 1973 y creó un manejador de base de datos usable llamado Ingres en 1974, que se conoció como el hermano menor del manejador de base de datos para entornos de mainframe de IBM, llamado System-R. Comercializó Ingres un tiempo y volvió al mundo académico donde se enfocó en un proyecto académico que se denominó Postgres (post-ingres), en 1985.

Paralelamente, Larry Ellison, inspirado en las ideas de Codd, desarrolló para la CIA un manejador de base de datos llamado Oracle (el Oráculo). En 1977 fundó una compañía llamada Software Development Laboratories, que en 1979 pasó a llamarse Relational Software Inc. (comenzando a comercializar Oracle, el primer RDBMS comercial con soporte para consultas SQL) y en 1983 Oracle Corporation.

En 1976, Peter Chen había publicado uno de los trabajos más importantes para la disciplina del diseño de base de datos, donde presenta el Modelo Entidad-Relación. Este modelo fue aceptado de facto como modelo para diseñar los esquemas lógicos de las bases de datos.

Durante los 80 se fundó Informix, y se intensificó la comercialización de sistemas manejadores de base de datos, así como la competencia. IBM lanzó comercialmente DB2 en 1983. Se estandarizó SQL (Structured Query Language) como el lenguaje de consultas de todos los

manejadores relacionales (convertido en un estándar ANSI en 1986, y ratificado por ISO en 1987), y el modelo Entidad-Relación para el diseño conceptual de las bases de datos.

En 1985, Codd publicó otro trabajo con una lista de 12 reglas que definen una base relacional ideal, dando una guía para el diseño de las bases de datos relacionales. Los proveedores han tratado de implementar manejadores de base de datos que conformen con las reglas de Codd, aunque conformar con todas es muy difícil.

Hasta los 90, IBM dominaba el mercado de las bases de datos en ambientes de mainframe con su producto DB2, mientras que Oracle e Informix dominaban (y competían por) el mercado en entornos UNIX y Microsoft Windows. Más tarde Microsoft también entraría en escena con su producto SQL Server, sólo para plataformas Microsoft Windows.

En 1994, el proyecto Postgres de Stonebraker tenía un manejador estable, con soporte de SQL, que llamaron Postgres95, nombre que duró poco ya que en 1996 se pasó a llamar PostgreSQL. Stonebraker fundó Illustra Information Technologies para comercializar una versión de PostgreSQL. La compañía fue comprada por Informix en 1997 y se produjo una transferencia tecnológica hacia la nueva versión de Informix (y Stonebraker se convirtió en el Chief Technology Officer de Informix).

A finales de los 90, el crecimiento exponencial de Internet generó la necesidad de un cambio en la forma de conectarse a las bases de datos y aparecieron tecnologías como ODBC (Open DataBase Conectivity), ASP (Active Server Pages), JSP (Java Server Pages), JDBC (Java DataBase Conectivity) y EJB (Enterprise Java Beans). Por estos años, Michael Stonebraker fundó Cohera Software, una empresa enfocada en el desarrollo de una base de datos federada. En 2001 Cohera Software fue comprada por PeopleSoft, que luego fue comprada por Oracle Corporation.

Aparecieron también algunos manejadores de base de datos gratuitos, como MySQL que ganó terreno como manejador de bases de datos para soporte de sitios Web (sobre todo desarrollados en PHP), y PostgreSQL que se liberó con licencia BSD. Es común hoy en día hablar de programadores LAMP (Linux, Apache, MySQL y PHP) en referencia al dominio de estas tecnologías gratuitas.

Cerca del año 2000, Microsoft había ganado mucho terreno con sus sistemas operativos Windows para servidores y su manejador de base de datos SQL Server; mientras que Oracle parecía haber vencido a Informix en entornos UNIX. En 2001 IBM compró Informix y comenzó un proceso de transferencia tecnológica hacia DB2. El segmento de mercado dominado por IBM también creció, al contar con todos los usuarios de DB2 más los de Informix.

Actualmente IBM continúa dominando el mercado en entornos de mainframe con DB2, IBM (con DB2 e Informix) compite con Oracle en entornos UNIX (y Linux) y los dos compiten con Microsoft (SQL Server) en ambientes Windows. Todos ellos compiten también con bases gratuitas como PostgreSQL y MySQL, y con versiones comerciales sobre estas gratuitas. También hay otros competidores de menor peso, como Teradata o Sybase.

A modo de respuesta a la oferta de manejadores gratuitos, los grandes de las bases de datos propietarias han liberado ediciones gratuitas de sus manejadores (Oracle Express, DB2 Express-C, SQL Server Express), con características limitadas (funcionalidades y hardware soportado).

Queda claro que las bases de datos relacionales desplazaron del mercado a las bases de datos jerárquicas, aunque en ambientes de mainframe continúan existiendo bases de datos jerárquicas como Adabas.

Como era de esperarse, las líneas de investigación no terminan con el modelo relacional, ya que el advenimiento de los lenguajes de programación orientados a objetos provocó el llamado "Object-Relational Impedance Mismatch", que básicamente significa una dificultad para manejar dos mundos diferentes. Se han diseñado dos soluciones a este problema, una es cambiar de paradigma y utilizar bases de datos orientadas a objetos (OODBMS) o relacionales

con soporte para objetos (ORDBMS), y la otra utilizar alguna especie de adaptador entre los dos mundos, manteniendo el motor puramente relacional y aprovechando los años de madurez de esta tecnología.

Las bases de datos orientadas a objetos han sido poco más que una curiosidad académica, y ninguno de los líderes del mercado se han vuelto en esta dirección. Por otro lado, hay toda una disciplina centrada en proveer un mapping entre los objetos y el modelo relacional (ORM: Object-Relational Mapping); se pueden mencionar estándares de Java como JPA (Java Persistence API) y JDO (Java Data Objects), y frameworks como Hibernate, cuyas funcionalidades se han integrado al propio lenguaje Java en la versión 3 de los Enterprise Java Beans (EJB3).

Stonebraker ha sido el que ha intentado dotar de capacidades de orientación a objetos a su motor (PostgreSQL), y que ha reconocido que el lenguaje SQL tiene dificultades inherentes para el manejo de objetos. Por otro lado, un trabajo de Date y Darwen (ambos vinculados a IBM), llamado "The Third Manifesto", esbozaba en 1995 una propuesta de trabajo futuro sobre los manejadores de bases de datos relacionales para evitar el Object-Relational Impedance Mismatch, pero manteniendo al SQL como el lenguaje universal para bases de datos.

Otras áreas actuales de investigación y desarrollo son la integración de XML y el soporte de XQuery (el lenguaje que permite manipular XML), el manejo de bases de datos muy grandes (VLDB) que llegan a tener varios TB de datos, la federación de datos, el soporte de datos espaciales y las técnicas asociadas a la gestión de DataWarehouses como las funciones OLAP o la minería de datos. Se podría mencionar también la continuidad en el intento de conformar con las 12 reglas de Codd, aunque este es un interés más académico que comercial, y la tecnología siempre avanza más en el sentido que pretenden los que financian las investigaciones.

Por todo esto, el modelo relacional parece tener para muchos años más en el plano de las bases de datos, y es por esto que el énfasis de este curso estará en este modelo. Pero antes de pasar a estudiar el modelo relacional será bueno recordar en las próximas clases algunos cuantos conceptos básicos de lógica y matemáticas que serán muy útiles.