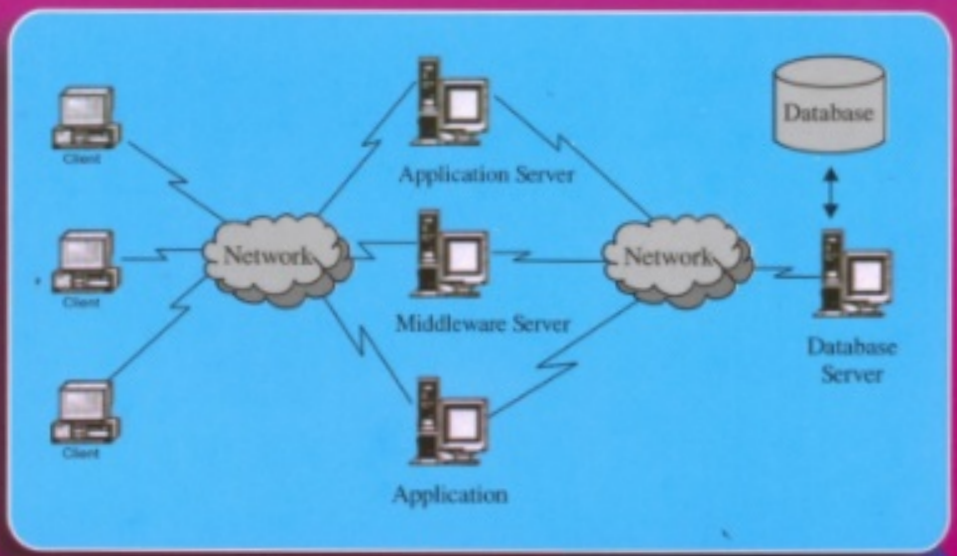
NGUYỄN VĂN HUÂN - PHẠM VIỆT BÌNH

GIÁO TRÌNH
HỆ CƠ SỞ DỮ LIỆU
PHÂN TÁN & SUY DIỄN

Lý thuyết và thực hành



NHÀ XUẤT BẢN KHOA HỌC VÀ KỸ THUẬT

NGUYỄN VĂN HUÂN - PHẠM VIỆT BÌNH

GIÁO TRÌNH

HỆ CƠ SỞ DỮ LIỆU PHÂN TÁN VÀ SUY DIỄN

NHÀ XUẤT BẢN KHOA HỌC VÀ KỸ THUẬT
Hà Nội - 2009

MỤC LỤC

CHƯƠNG I
GIỚI THIỆU CHUNG VỀ CƠ SỞ DỮ LIỆU

1.1. KHÁI NIỆM CƠ BẢN VỀ CÁC HỆ CƠ SỞ DỮ LIỆU

Trong chương này trình bày những khái niệm cơ bản về các hệ cơ sở dữ liệu do E.F Codd đề xuất. Những khái niệm này bao gồm mục tiêu của một hệ cơ sở dữ liệu. Sự cần thiết phải tổ chức dữ liệu dưới dạng cơ sở dữ liệu. Tính độc lập của dữ liệu thể hiện mô hình kiến trúc 3 mức. Vì vậy có thể nói cơ sở dữ liệu phản ánh tính trung thực, khách quan của thế giới dữ liệu. Không dư thừa thông tin và cũng không thiếu thông tin. Nội dung của chương bao gồm các phần:

- Cơ sở dữ liệu là gì;
- Sự cần thiết của các hệ cơ sở dữ liệu;
- Mô hình kiến trúc 3 mức cơ sở dữ liệu;
- Mục tiêu của các hệ cơ sở dữ liệu;
- Hệ quản trị CSDL & người quản trị CSDL;
- Tổ chức lưu trữ dữ liệu;
- Các mô hình truy xuất .

1.1.1. Cơ sở dữ liệu là gì

Cơ sở dữ liệu là một bộ sưu tập rất lớn về các loại dữ liệu tác nghiệp, bao gồm các loại dữ liệu âm thanh, tiếng nói, chữ viết, văn bản, đồ hoạ, hình ảnh tĩnh hay hình ảnh động... được mã hoá dưới dạng các chuỗi bit và được lưu trữ dưới dạng File dữ liệu trong các bộ nhớ của máy tính. Cấu trúc lưu trữ dữ liệu tuân theo các quy tắc dựa trên lý thuyết toán học. Cơ sở dữ liệu phản ánh trung thực thế giới dữ liệu hiện thực khách quan.

Cơ sở dữ liệu là tài nguyên thông tin dùng chung cho nhiều người: Cơ sở dữ liệu (CSDL) là tài nguyên thông tin chung cho nhiều người cùng sử dụng. Bất kỳ người sử dụng nào trên mạng máy tính, tại các thiết bị đầu cuối, về nguyên tắc có quyền truy nhập khai thác toàn bộ hay một phần dữ liệu theo chế độ trực tuyến hay tương tác mà không phụ thuộc vào vị trí địa lý của người sử dụng với các tài nguyên đó.

Cơ sở dữ liệu được các hệ ứng dụng khai thác bằng ngôn ngữ con dữ liệu hoặc bằng các chương trình ứng dụng để xử lý, tìm kiếm, tra cứu, sửa đổi, bổ sung hay loại bỏ dữ liệu. Tìm kiếm và tra cứu thông tin là một trong những chức năng quan trọng và phổ biến nhất của các dịch vụ cơ sở dữ liệu. Hệ quản trị CSDL – HQTCSDL (DataBase Management System - DBMS) là phần mềm điều khiển các chiến lược truy nhập CSDL. Khi người sử dụng đưa ra yêu cầu truy nhập bằng một ngôn ngữ con dữ liệu nào đó, HQTCSDL tiếp nhận và thực hiện các thao tác trên CSDL lưu trữ.

Đối tượng nghiên cứu của CSDL là các thực thể và mối quan hệ giữa các thực thể. Thực thể và mối quan hệ giữa các thực thể là hai đối tượng khác nhau về căn bản. Mối quan hệ giữa các thực thể cũng là một loại thực thể đặc biệt. Trong cách tiếp cận CSDL quan hệ, người ta dựa trên cơ sở lý thuyết đại số quan hệ để xây dựng các quan hệ chuẩn, khi kết nối không tổn thất thông tin và khi biểu diễn dữ liệu là duy nhất. Dữ liệu được lưu trữ trong bộ nhớ của máy tính không những phải tính đến yếu tố về tối ưu không gian lưu trữ, mà phải đảm bảo tính khách quan, trung thực của dữ liệu hiện thực. Nghĩa là phải đảm bảo tính nhất quán của dữ liệu và giữ được sự toàn vẹn của dữ liệu.

1.1.2. Sự cần thiết của các hệ cơ sở dữ liệu

Tổ chức lưu trữ dữ liệu theo lý thuyết cơ sở dữ liệu có những ưu điểm:

Giảm bớt dư thừa dữ liệu trong lưu trữ: Trong các ứng dụng lập trình truyền thống, phương pháp tổ chức lưu trữ dữ liệu vừa tốn kém, lãng phí bộ nhớ và các thiết bị lưu trữ, vừa dư thừa thông tin lưu trữ. Nhiều chương trình ứng dụng khác nhau cùng xử lý trên các dữ liệu như nhau, dẫn đến sự dư thừa đáng kể về dữ liệu. Ví dụ trong các bài toán nghiệp vụ quản lý "Cước thuê bao điện thoại" và "Doanh thu & sản lượng", tương ứng với mỗi một chương trình là một hay nhiều tệp dữ liệu được lưu trữ riêng biệt, độc lập với nhau. Trong cả 2 chương trình cùng xử lý một số thuộc tính của một cuộc đàm thoại như "số máy gọi đi", "số máy gọi đến", "hướng cuộc gọi", "thời gian bắt đầu" và "thời gian kết thúc"... Nhiều thuộc tính được mô tả và lưu trữ nhiều lần độc lập với nhau. Nếu tổ chức lưu trữ theo lý thuyết CSDL thì có thể hợp nhất các tệp lưu trữ của các bài toán trên, các chương trình ứng dụng có thể cùng chia sẻ tài nguyên trên cùng một hệ CSDL.

Tổ chức lưu trữ dữ liệu theo lý thuyết CSDL sẽ tránh được sự không nhất quán trong lưu trữ dữ liệu và bảo đảm được tính toàn vẹn của dữ liệu: Nếu một thuộc tính được mô tả trong nhiều tệp dữ liệu khác nhau và lặp lại nhiều lần trong các bản ghi, khi thực hiện việc cập nhật, sửa đổi, bổ sung sẽ không sửa hết nội dung các mục đó. Nếu dữ liệu càng nhiều thì sự sai sót khi cập nhật, bổ sung càng lớn. Khả năng xuất hiện mâu thuẫn, không nhất quán thông tin càng nhiều, dẫn đến không nhất quán dữ liệu trong lưu trữ. Tất yếu kéo theo sự dị thường thông tin, thừa, thiếu và mâu thuẫn thông tin.

Thông thường, trong một thực thể, giữa các thuộc tính có mối quan hệ ràng buộc lẫn nhau, tác động ảnh hưởng lẫn nhau. Cước của một cuộc đàm thoại phụ thuộc vào khoảng cách và thời gian cuộc gọi, tức là phụ thuộc hàm vào các thuộc tính máy gọi đi, máy gọi đến, thời gian bắt đầu và thời gian kết thúc cuộc gọi. Các trình ứng dụng khác nhau cùng xử lý cước đàm thoại trên các thực thể lưu trữ tương ứng khác nhau chưa hẳn cho cùng một kết quả về sản lượng phút và doanh thu. Điều này lý giải tại sao trong một doanh nghiệp, cùng xử lý trên các chỉ tiêu quản lý mà số liệu báo cáo của các phòng ban, các công ty con lại cho các kết quả khác nhau, thậm chí còn trái ngược nhau. Như vậy, có thể khẳng định, nếu dữ liệu không tổ chức theo lý thuyết cơ sở dữ liệu, tất yếu không thể phản ánh thế giới hiện thực dữ liệu, không phản ánh đúng bản chất vận động của dữ liệu.

Sự không nhất quán dữ liệu trong lưu trữ làm cho dữ liệu mất đi tính toàn vẹn của nó. Tính toàn vẹn dữ liệu đảm bảo cho sự lưu trữ dữ liệu luôn luôn đúng. Không thể có mã vùng ngoài quy định của cơ quan quản lý, hoặc ngày sinh của một nhân viên không thể xảy ra sau ngày tốt nghiệp ra trường của nhân viên đó...

Tổ chức lưu trữ dữ liệu theo lý thuyết CSDL có thể triển khai đồng thời nhiều ứng dụng trên cùng một CSDL: Điều này có nghĩa là các ứng dụng không chỉ chia sẻ chung tài nguyên dữ liệu mà còn trên cùng một CSDL có thể triển khai đồng thời nhiều ứng dụng khác nhau tại các thiết bị đầu cuối khác nhau.

Tổ chức dữ liệu theo lý thuyết cơ sở dữ liệu sẽ thống nhất các tiêu chuẩn, thủ tục và các biện pháp bảo vệ, an toàn dữ liệu: Các hệ CSDL sẽ được quản lý tập trung bởi một người hay một nhóm người quản trị CSDL, bằng các hệ quản trị CSDL. Người quản trị CSDL có thể áp dụng thống nhất các tiêu chuẩn, quy định, thủ tục chung như quy định thống nhất về mẫu biểu báo cáo, thời gian bổ sung, cập nhật dữ liệu. Điều này làm dễ dàng cho công việc bảo trì dữ liệu. Người quản trị CSDL có thể bảo đảm việc truy nhập tới CSDL, có thể kiểm tra, kiểm soát các quyền truy nhập của người sử dụng. Ngăn chặn các truy nhập trái phép, sai quy định từ trong ra hoặc từ ngoài vào...

1.1.3. Mô hình kiến trúc tổng quát cơ sở dữ liệu 3 mức

Mô hình kiến trúc 3 mức của hệ CSDL gồm: *Mức trong, mức mô hình dữ liệu (Mức quan niệm) và mức ngoài*. Giữa các mức tồn tại các ánh xạ quan niệm trong và ánh xạ quan niệm ngoài. Trung tâm của hệ thống là mức quan niệm, tức là mức mô hình dữ liệu. Ngoài ra còn có khái niệm người sử dụng, hệ quản trị CSDL và người quản trị CSDL.

Người sử dụng: Là những người tại thiết bị đầu cuối truy nhập vào các hệ CSDL theo chế độ trực tuyến hay tương tác bằng các chương trình ứng dụng hay bằng các ngôn ngữ con dữ liệu. Thường là các chuyên viên kỹ thuật tin học, có trình độ thành thạo biết lập trình và biết sử dụng ngôn ngữ con thao tác dữ liệu (SQL Server, Oracle...). Người sử dụng có thể truy nhập toàn bộ hay một phần CSDL mà họ quan tâm, phụ thuộc vào quyền truy nhập của họ. Cách nhìn CSDL của người sử dụng nói chung là trìu tượng. Họ nhìn CSDL bằng mô hình ngoài, gọi là mô hình con dữ liệu. Chẳng hạn người sử dụng là một nhân viên của phòng kế toán tài chính, chỉ nhìn thấy tập các xuất hiện kiểu bản ghi ngoài về doanh thu, sản lượng trong tháng, không thể nhìn thấy các xuất hiện kiểu bản ghi lưu trữ về các chỉ tiêu kỹ thuật của đường thông, mạng lưới...

Mô hình ngoài: Mô hình ngoài là nội dung thông tin của CSDL dưới cách nhìn của người sử dụng. Là nội dung thông tin của một phần dữ liệu tác nghiệp được một người hoặc một nhóm người sử dụng quan tâm. Nói cách khác, mô hình ngoài mô tả cách nhìn dữ liệu của người sử dụng và mỗi người sử dụng có cách nhìn dữ liệu khác nhau. Nhiều mô hình ngoài khác nhau có thể cùng tồn tại trong một hệ CSDL, nghĩa là có nhiều người sử dụng chia sẻ chung cùng một cơ sở dữ liệu. Hơn nữa, có thể mô hình ngoài quan hệ, mô hình ngoài phân cấp hay mô hình ngoài kiểu mạng cũng có thể tồn tại trong một cơ sở dữ liệu. Sơ đồ ngoài không làm "hiện" mà được nhúng vào trong logic một đơn tác có liên quan.

• Mô hình ngoài gồm nhiều xuất hiện kiểu bản ghi ngoài, nghĩa là mỗi một người sử dụng có một sơ đồ dữ liệu riêng, một khung nhìn dữ liệu riêng. Bản ghi ngoài của người sử dụng có thể khác với bản ghi lưu trữ và bản ghi quan niệm.

• Mô hình ngoài được xác định bởi một sơ đồ ngoài bao gồm các mô tả về kiểu bản ghi ngoài như tên các trường, kiểu dữ liệu các trường, độ rộng của trường...

• Ngôn ngữ con dữ liệu của người sử dụng thao tác trên các bản ghi ngoài.