|  |
| --- |
| 1. KHÁI NIỆM CHUNG VỀ CƠ SỞ DỮ LIỆU |
| Theo nhận thức chung nhất của nhiều độc giả, một **cơ sở dữ liệu** đơn giản là một bộ tập hợp các dữ liệu liên quan tới nhau. Định nghĩa này khá mơ hồ bởi vì nếu áp dụng định nghĩa thê này, chúng ta có thể xem một trang sách là một cơ sở dữ liệu bởi vì nó bao gồm các dữ liệu là những từ nằm trong tranh sách đó và rõ ràng các từ này có quan hệ với nhau vì chúng cùng mô tả nội dung một chủ đề cụ thể nào đó đang được thể hiện trong trang sách đó. Lưu ý rằng khái niệm “dữ liệu” trong một cơ sở dữ liệu có thể bao phủ một phạm vi rất rộng các đối tượng khác nhau từ các số, văn bản, đồ họa, video, v.v… Một định nghĩa cụ thể hơn nữa của một cơ sở dữ liệu bao gồm các đặc tính không tường minh được cân nhắc cùng nhau để định nghĩa một cơ sở dữ liệu. Chúng ta cùng xem xét cách nhìn nhận khái niệm cơ sở dữ liệu theo cách cụ thể này. Một cơ sở dữ liệu thể hiện các khía cạnh khác nhau của một thế giới thực. Sự trừu tượng của một thế giới thực thường được coi là một thế giới nhỏ hoặc vũ trụ của một vấn đề nào đó. Một cách khác, một cơ sở dữ liệu được coi là một bộ thu thập dữ liệu với các ý nghĩa gắn kết. Các dữ liệu ngẫu nhiên thường không thể coi là một cơ sở dữ liệu mặc dù chúng là những ngoại lệ. |
| 1.1. Các khái niệm cơ bản |
| Một cơ sở dữ liệu được thiết kế, xây dựng, lớn dần và được sử dụng cho một mục đích cụ thể nào đó. Nó sẽ có một tập các người sử dụng tiềm năng và được sử dụng cho các ứng dụng cụ thể ngay từ khi thiết kế ban bầu. Ví dụ một cơ sở dữ liệu quản lý thông tin của sinh viên trong một trường học, được dùng với mục đích quản lý các hoạt động chính của sinh viên trong trường bao gồm một số chức năng chính là quản lý điểm số các môn học, quản lý thi đua, được sử dụng bởi nhóm người dùng tiềm năng là sinh viên, các cán bộ quản lý và giáo viên trong trường… |
| 1.1.1. Các hệ thống cơ sở dữ liệu truyền thống |
| Các hệ thống cơ sở dữ liệu thương mại đầu tiên xuất hiện vào những năm của thập kỷ 60. Chúng phát triển từ hệ thống lưu trữ theo kiểu tệp truyền thống, theo kiểu tệp này thì các dữ liệu được lưu trữ trong các tệp tách rời nhau và được lưu trữ trong bộ lưu trữ vật lý. Các hệ thống tệp này có cung cấp các đặc tính (3) của hệ quản trị cơ sở dữ liệu được mô tả ở phần trên nhưng chúng không hoặc cung cấp rất ít các tính năng (4). Hơn nữa, các hệ thống tệp này không cung cấp các chức năng hỗ trợ trực tiếp cho các tính năng của mục (2) ở trên ví dụ như chúng không hỗ trợ các ngôn ngữ truy vấn. Hệ thống này cũng không hỗ trợ trực tiếp chức năng ở mục (1) ở trên, việc hỗ trợ cho các lược đồ quan hệ rất hạn chế, chỉ cho phép tạo cấu trúc thư mục cho các tệp dữ liệu mà không cho phép thay đổi hay tạo mới cấu trúc của các tệp. Một vài hệ thống cơ sở dữ liệu ban đầu quan trọng hơn là những hệ thống trong đó dữ liệu được cấu tạo bởi nhiều mục nhỏ và nhiều truy vấn hoặc thay đổi có thể thực hiện được. Ví dụ như các hệ thống đặt vé máy bay hay các hệ thống ngân hàng. Nói đến sự phát triển vượt bậc của các hệ thống cơ sở dữ liệu phải nhắc đến một bài báo nổi tiếng được viết bởi Codd năm 1970, một bài báo có ảnh hưởng rất lớn tới sự thay đổi các hệ thống cơ sở dữ liệu (tham khảo tới bài báo Codd, E.F., “A relational model for large shared data banks”, Communications of ACM, **13**:6, pp. 377-387). Trong bài báo này, Codd đề xuất rằng các hệ thống cơ sở  dữ liệu nên đưa ra cho người sử dụng một khung nhìn về dữ liệu trong đó dữ liệu được tổ chức dưới dạng các bảng được gọi là các quan hệ. Bên trong sự mô tả dữ liệu theo kiểu này, một cấu trúc dữ liệu phức tạp được thiết lập cho phép các truy vấn của người sử dụng được đáp ứng nhanh chóng. Nhưng không giống như những người sử dụng các hệ thống cơ sở dữ liệu trước đây, người dùng của một hệ thống quan hệ không cần quan tâm tới cấu trúc lưu trữ của dữ liệu. Các câu truy vấn sau đó có thể được thể hiện trong một ngôn ngữ bậc cao, là loại ngôn ngữ làm tăng hiệu suất đáng kể cho những người lập trình cơ sở dữ liệu.  Các hệ thống cơ sở dữ liệu ban đầu có kích cỡ rất lớn, một loại trong số chúng là hệ quản trị cơ sở dữ liệu. Ban đầu hệ quản trị cơ sở dữ liệu rát lớn, có giá thành cao và chạy trên các máy tính mainframe lớn.  Kích cỡ bộ lưu trữ hàng gigabytes dữ liệu trước đây là rất lớn nên cần các máy tính lớn. Ngày nay, một gigabyte có thể được lưu trữ trên một đĩa đơn do công nghệ ngày càng phát triển, vì vậy hệ quản trị cơ sở dữ liệu có thể chạy trên một máy tính cá nhân là hoàn toàn khả thi. Hệ thống ngày càng nhỏ dần theo thời gian do công nghệ điện tử càng phát triển. Ngày nay, hệ quản trị cơ sở dữ liệu dựa trên mô hình dữ liệu bắt đầu xuất hiện như một công cụ chung cho các ứng dụng của máy tính cũng như việc các bảng tính và các bộ xử lý văn bản đã trở thành công cụ chung trước đây. Mặt khác, một gigabyte ngày nay không còn được coi là dữ liệu có kích cỡ lớn nữa. Các hệ cơ sở dữ liệu lớn phải chứa hàng trăm gigabytes hoặc nhiều hơn. Khi bộ nhớ lưu trữ trở nên rẻ hơn, con người thường tìm thấy các lý do mới để lưu trữ nhiều dữ liệu hơn. Chẳng hạn, một chuỗi các cửa hàng bán lẻ thường lưu trữ tới terabytes (1 terabytes = 1000 gigabytes hoặc 1012 bytes) thông tin để lưu lại lịch sử của mỗi giao dịch mua bán trong một khoảng thời gian rất dài. Dữ liệu thì không phải chỉ ở dạng văn bản và số như trước kia, mà nay có nhiều dạng mới như dạng âm thanh, hình ảnh thường chiếm không gian lưu trữ rất lớn ví dụ như một giờ của video sẽ chiếm một gigabyte. Hay các cơ sở dữ liệu lưu trữ các hình ảnh vệ tinh sẽ chiếm nhiều petabytes dữ liệu trong đó 1 petabyte=1000 gigabytes hay 1015 bytes. Như vậy một xu hướng hiện này là dữ liệu ngày càng lớn.  Để quản lý được các cơ sở dữ liệu lớn như vậy đòi hỏi nhiều công nghệ hiện đại   * Các cơ sở dữ liệu hiện đại của các kích cỡ gần đây nhất được lưu trữ trên một mảng các đĩa (các thiết bị lưu trữ thứ cấp) * Các cơ sở dữ liệu hầu như không bao giờ cho rằng “dữ liệu” sẽ vuwafvowis bộ nhớ trong. Các hệ thống cũ cũ hơn thường chỉ có các thiết bị lưu trữ thứ cấp dưới dạng các đĩa từ (công nghệ tương tự)   Hai hướng mới cho phép các hệ thống cơ sở dữ liệu có thể quản lý được khối lượng dữ liệu lớn hơn một cách nhanh hơn là:   1. Lưu trữ mức độ cấp 3: các cơ sở dữ liệu lớn nhất hiện này đòi hỏi nhiều hơn chỉ lưu trữ trên các đĩa (cấp 2). Các thiết bị cấp 3 có xu hướng lưu trữ theo đơn vị terabyte và có thời gian truy nhập dài hơn thời gian truy nhập của đĩa từ truyền thống. Thời gian truy nhập của một đĩa truyền thống là nằm trong khoảng 10-20 m giây. Trong khi đó của thiết bị mới này sẽ mất vài giây. Các thiết bị này liên quan tới việc chuyển các đối tượng mà trên đó dữ liệu được lưu trữ tới một thiết bị đọc nào đó thông qua một dạng giám sát bằng robot nào đó. Việc sử dụng đĩa CDs như một phương tiện lưu trữ mức độ này. 2. Tính toán song song: Khả năng lưu trữ khối lượng dữ liệu khổng lồ là rất quan trọng nhưng nó sẽ ít được sử dụng nếu như chúng ta không thể truy nhập vào khối lượng lớn dữ liệu đó một cách nhanh chóng. Các cơ sở dữ liệu rất lớn đòi hỏi bộ cải thiện tốc đọ. Việc cải thiện tốc độ được thực hiện bằng nhiều cách trong cơ sở dữ liệu hiện đại ngày nay bao gồm:    * Các cấu trúc chỉ mục    * Cơ chế song song hóa- liên quan tới cả song song hóa bộ vi xử lý cũng như song song hóa bản thân dữ liệu. Trong một phạm vi nào đó, các hệ thống cơ sở dữ liệu phân tán cũng có thể được cho vào như một bộ cải thiện tốc độ mặc dù theo một cách thức hơi khác, chúng ta sẽ xem xét vấn đề này trong các bài giảng sau. |
| 1.1.2. Cơ sở dữ liệu và hệ quản trị cơ sở dữ liệu |
| 1.1.2.2. Hệ quản trị cơ sở dữ liệu |
| Một cơ sở dữ liệu được quản lý bởi một hệ quản trị cơ sở dữ liệu, thường được tham khảo tới như một hệ thống cơ sở dữ liệu. Một hệ quản trị cơ sở dữ liệu được cho là sẽ phải cung cấp những tính năng quan trọng sau đây:   1. Cho phép người sử dụng tạo ra một cơ sở dữ liệu mới. Việc này sẽ được thực hiện thông qua một ngôn ngữ định nghĩa dữ liệu (Data Definition Languages-DDLs). 2. Cho phép người sử dụng truy vấn cơ sở dữ liệu thông qua ngôn ngữ thao tác dữ liệu (Data Manipulaton Languages-DMLs). 3. Hỗ trợ việc lưu trữ một khối lượng rất lớn dữ liệu mà không gây mất mát và tổn thất thông tin. Kích cỡ điển hình là từ nhiều gigabytes trở lên và lưu trữ chúng hiệu quả trong một khoảng thời gian rất dài. Đương nhiên để lưu trữ tốt trong khoảng thời gian dài đó thì cần phải duy trì , cập nhật thông tin tốt và hiệu quả. Đồng thời, duy trì tính bảo mật và tính toàn vẹn dữ liệu trong các xử lý được thực hiện trong hệ thống. 4. Kiểm soát truy nhập dữ liệu từ nhiều người sử dụng cùng một lúc. |
| 1.1.2.2.1. Các thành phần của hệ quản trị cơ sở dữ liệu |
| Cơ sở dữ liệu được lưu trữ và siêu dữ liệu: Cơ sở dữ liệu được lưu trữ nằm trên các thiết bị cấp 3 hoặc thứ cấp (cấp 2). Trong một vài thời điểm, một số phần của cơ sở dữ liệu sẽ được lưu trữ thêm trên các bộ lưu trữ đệm nhưng chúng ta sẽ bỏ qua vấn đến này trong phạm vi bài giảng này.   |  | | --- | |  | |  |   Siêu dữ liệu là dữ liệu về dữ liệu. Trong trường hợp này siêu dữ liệu là các mô tả về các thành phần dữ liệu của cơ sở dữ liệu. Vị trí tương đối của các trường trong một bản ghi, các thông tin về lược đồ, thông tin về chỉ mục và nhiều thứ khác nữa. Với một cơ sở dữ liệu nào đó, một hệ quản trị cơ sở dữ liệu có thể duy trì nhiều bộ chỉ mục khác nhau được thiết kế để cung cấp truy nhập nhanh tới dữ liệu ngẫu nhiên. Hầu hết các chỉ mục này được thể hiện như dạng cây nhị phân trong các cơ sở dữ liệu hiện đại. Cây nhị phân có xu hướng ngắn và béo khiến việc truy nhập nhanh từ gốc đến lá. |
| 1.1.2.2.2. Kiến trúc hệ quản trị cơ sở dữ liệu |
| Bộ quản lý lưu trữ: trong một hệ thống cơ sở dữ liệu đơn giản, bộ quản lý lưu trữ chỉ là đơn giản là một hệ thống tệp của một hệ điều hành nào đó. Trong các hệ thống lớn hơn, để hiệu quả, các kho lưu trữ của hệ quản trị cơ sở dữ liệu được quản lý trên đĩa một cách trực tiếp. Bộ quản lý lưu trữ bao gồm hai thành phần cơ bản (1) quản lý vùng đệm và (2) quản lý tệp:   * Quản lý tệp làm nhiệm vụ theo dõi vị trí của các tệp trên các đĩa và lấy ra được các khối hoặc khối chứa một tệp yêu cầu từ bộ quản lý đệm. Các đĩa thường bị chặn vào các vùng có không gian liên tục kích cỡ từ 212 đến 214 bytes (khoảng 4000  đến 16,000 bytes/1 khối dữ liệu). * Bộ quản lý đệm quản lý bộ nhớ chính. Các khối dữ liệu được lấy ra từ các đĩa thông qua bộ quản lý tệp và lựa chọn một trang trong bộ nhớ chính để lưu trữ chứ không chặn lại. Thuật toán thiết lập trang sẽ quyết định một trang sẽ tồn tại bao lâu trong bộ nhớ chính. Tuy nhiên, bộ quản lý các giao dịch có thể ép một trạng trong bộ nhớ chính quay trở về đĩa (chúng ta sẽ xem xét những chi tiết của vấn đề này trong các bài sau).  |  | | --- | |  | |  |   Bộ quản lý truy vấn làm nhiệm vụ chuyển một truy vấn hoặc thao tác cơ sở dữ liệu, đang được thể hiện bằng một ngôn ngữ bậc cao (ví dụ SQL) thành một chuỗi các yêu cầu cho các dữ liệu được lưu trữ như các bộ cụ thể của một quan hệ hoặc các phần của một chỉ mục của một quan hề. Phần khó khăn nhất của việc xử lý truy vấn thường là quá trình tối ưu hóa truy vấn, cái liên quan tới việc thiết lập một dạng chiến lược tốt để thực thi truy vấn. Chúng ta sẽ bàn luận tới tối ưu hóa truy vấn một cách chi tiết hơn trong những phần sau của bài giảng. Bộ quản lý giao dịch: tồn tại một số đảm bảo cho một hệ quản trị cơ sở dữ liệu phải có khi thực hiện các phép toán trên một cơ sở dữ liệu. Những đảm bảo này thường được nhắc tới như các thuộc tính ACID được mô tả như sau:   * Tính nguyên vẹn: được thể hiện theo nguyên tắc hoặc tất cả công việc của một giao dịch được thực hiện hoặc không công việc nào của nó được thực hiện * Tính đồng nhất: có nghĩa là dữ liệu không thể ở một trạng thái không đồng nhất * Tính biệt lập: có nghĩa là các giao dịch đồng thời phải được tách riêng một cách biệt lập khỏi nhau trên cả phương diện ảnh hưởng và tính hiển thị * Tính duy trì: những thay đổi tới cơ sở dữ liệu gây ra bởi một giao dịch không được mất đi thậm chí hệ thống bị hỏng ngay sau khi giao dịch hoàn thành. |
| 1.1.2.2.3. Các ưu điểm của hệ quản trị cơ sở dữ liệu |
| |  |  | | --- | --- | | **Muc đích** | **Ý nghĩa** | | Kiểm soát sự dư thừa dữ liệu | Kinh tế khi tăng khối lượng dữ liệu | | Đảm bảo dữ liệu đồng nhất | Tạo cân bằng cho các yêu cầu bị xung đột | | Cung cấp thêm thông tin từ dữ liệu | Cải thiện tính truy nhập của dữ liệu | | Hỗ trợ sự sẵn có của khối lượng lớn dữ liệu | Cải thiện năng suất thực hiện của hệ thống | | Hỗ trợ chia sẻ dữ liệu | Cải thiện sự bảo dường | | Cải thiện sự toàn vẹn dữ liệu | Tăng xử lý đồng thời | | Cải thiện bảo mật dữ liệu | Cải thiện việc sao lưu và phục hồi | | Chuẩn hóa thông tin và dữ liệu | Cải thiện tính đáp ứng tới các truy vấn |   Bên cạnh những ưu điểm của các hệ quản trị cơ sở dữ liệu, các hệ thống này cũng có những nhược điểm như:   * Tính phức tạp * Kích cỡ lớn * Tốn chi phí mua và bảo trì * Thêm chi phí cho các phần cứng hỗ trợ * Chi phí chuyển đổi hệ thống * Hạn chế về năng suất thực hiện (cho một số trường hợp cụ thể) * Ảnh hưởng rất lớn đối với các trường hợp hỏng |
| 1.1.2.3. Khái niệm dữ liệu và thông tin |
| Xét một ví dụ về dữ liệu về các con số như sau:            0     11,500            5     12,300           10    12,800           15    10,455            20   12,200            25   13,900            30   14,220 Thể hiện dưới dạng “thô” như trên, dữ liệu này có rất ít ý nghĩa. Trong trường hợp này, nó đơn giản giống như một cặp danh sách các số nguyên. Ngữ cảnh để làm nền cho dữ liệu không có. Bằng việc xử lý dữ liệu, chúng ta chuyển đổi nó thành một dạng có ý nghĩa hơn. Trong ví dụ này quá trình xử lý bao gồm chủ yếu là đặt dữ liệu vào ngữ cảnh (cái thường được làm bằng cách thêm dữ liệu vào. Mặc dù những dữ liệu thêm vào này thực sự là siêu dữ liệu). Bây giờ dữ liệu bắt đầu có ý nghĩa hơn như sau: Thông tin: Dữ liệu RPM: Roebling Road 10/4/2003- Yamaha Heavy Vòng 12:  thời gian        rpm                        0          11,500                        5          12,300                          10          12,800                      15          10,455                         20          12,200                      25          13,900                         30          14,220 Với cùng dữ liệu được mô tả ở trên, xem xét cách xử lý sau đây chuyển đổi dữ liệu đó sang dạng đồ thị   |  | | --- | |  | |  | |
| 1.1.2.4. Dữ liệu phái sinh và dữ liệu vật lý |
| Một cơ sở dữ liệu sẽ chứa rất nhiều dữ liệu  vì nó mô hình hóa toàn bộ các tính năng và đặc điểm của một công ty đối tượng được thể hiện trong cơ sở dữ liệu này. Ví dụ, xem xét một cơ sở dữ liệu sinh viên của học viện công nghệ bưu chính viễn thông, chúng ta sẽ biểu diễn tên, số chứng minh thư, ngành học của mỗi sinh viên cùng với tập hợp những lớp học và điểm số tương ứng mà mỗi sinh viên đó tham gia. Tại một thời điểm nào đó trong quá trình hoạt động của hệ thống, một người có đủ thẩm quyền truy nhập sẽ nhập dữ liệu của sinh viên vào cơ sở dữ liệu theo một cách thức nào đó, hoặc là tự động hoặc là cách thủ công bằng tay.  Nếu chúng ta giả sử rằng cơ sở dữ liệu này cần lưu trữ điểm trung bình của mỗi sinh viên thì giá trị của các điểm trung bình sẽ xuất phát từ đâu? Liệu có cần nhập dữ liệu này vào từ người sử dụng không? Thường thì chúng ta không cần thiết phải làm như vậy, mà hệ quản trị cơ sở dữ liệu chỉ cần tính toán điểm trung bình từ các điểm số khác (thực ra là chỉ cần một chương trình phần mềm đơn giản chạy trên nền hệ quản trị là có thể làm được việc này, chúng ta sẽ bàn luận về vấn đề này trong phần sau của bài giảng). Vì vậy, điểm trung bình của một sinh viên sẽ được suy ra từ những dữ liệu khác liên quan tới sinh viên này. Nếu dữ liệu dùng để tính giá trị trung bình này thay đổi giá trị theo một cách thức nào đó thì giá trị của điểm số trung bình cũng thay đổi theo. Phụ thuộc vào mức độ phức tạp của tầng ứng dụng và của hệ quản trị cơ sở dữ liệu, khối dữ liệu phái sinh (được tính toán) từ những dữ liệu khác nằm trong cơ sở dữ liệu có thể lớn hơn rất nhiều khối lượng của dữ liệu được nhập vào trong thực tế (hay gọi là dữ liệu vật lý). Một thực tế khác là những cân nhắc khi nào và liệu dữ liệu phái sinh này có trở thành dữ liệu vật lý bởi vì không có hạn chế cho việc dữ liệu phái sinh sẽ trở thành thành viên chính thức được nhập vào cơ sở dữ liệu. |
| 1.2. Thiết kế cơ sở dữ liệu |
| 1.2.1. Sự cần thiết của việc thiết kế cơ sở dữ liệu |
| Đối với các hệ thống được xây dựng để phục vụ nhu cầu của người sử dụng, các hoạt động thiết kế cơ sở dữ liệu là cần thiết. Một cơ sở dữ liệu được thiết kế không cẩn thận sẽ tạo ra lỗi và lỗi này có thể dẫn tới việc đưa ra các kết luận không đúng đắn, dẫn tới những thiệt hại nặng nề cho tổ chức. Mặt khác, một cơ sở dữ liệu được thiết kế tốt sẽ sinh ra một hệ thống làm việc hiệu quả cung cấp các thông tin một cách chính xác hỗ trợ cho quá trình đưa ra quyết định đúng đắn dẫn tới thành công. |
| 1.2.2. Các vai trò cần thiết trong một môi trường cơ sở dữ liệu |
| Vai trò người quản trị dữ liệu(DA): là người có trách nhiệm cho việc quản lý các tài nguyên dữ liệu bao gồm việc lập kế hoạch cho cơ sở dữ liệu, phát triển và duy trì các chuẩn hóa và thủ tục, thiết kế cơ sở dữ liệu mức khái niệm và mức logic. Vai trò quản trị hệ thống cơ sở dữ liệu (DBA): là người chịu trách nhiệm với việc lưu trữ vật lý của cơ sở dữ liệu bao gồm thiết kế cơ sở dữ liệ vật lý và cài đặt, kiểm soát bảo mật và toàn vẹn của dữ liệu, bảo trò các hệ thống thao tác dữ liệu và đảm bảo công suất thỏa mãn người sử dụng cho các ứng dụng. vai trò DBA có thiên hướng liên quan tới kỹ thuật hơn là vai trò của DA. Vai trò người thiết kế cơ sở dữ liệu: trong các dự án thiết kế cơ sở dữ liệu có kích cỡ lớn, chúng ta có thể phân biệt hai loại thiết kế: người thiết kế cơ sở dữ liệu logic và người thiết kế cơ sở dữ liệu vật lý.   * Người thiết kế cơ sở dữ liệu logic liên quan tới việc xác định dữ liệu (các thực thể và các thuộc tính tương ứng), các mối quan hệ giữa dữ liệu, các ràng buộc trên dữ liệu được lưu trữ trong cơ sở dữ liệu. * Người thiết kế cơ sở dữ liệu vật lý phụ thuộc nhiều vào hệ quản trị cơ sở dữ liệu đích. Có nhiều cách để cài đặt cơ chế của cơ sở dữ liệu. Người thiết kế cơ sở dữ liệu vật lý phải nhận thức được toàn bộ các tính năng của hệ thống quản trị cơ sở dữ liệu.   Vai trò người phát triển ứng dụng: khi một cơ sở dữ liệu được cài đặt, các chương trình ứng dụng mà cung cấp các tính năng cần thiết cho người sử dụng cuối cần được xây dựng. Đây là trách nhiệm của người phát triển ứng dụng. Vai trò người sử dụng cuối: chính là khách hàng sử dụng cơ sở dữ liệu và có thể được phân thành hai nhóm người sử dụng dựa trên việc sử dụng hệ thống này như thế nào   * Nhóm thứ nhất được gọi là nhóm sử dụng không có kinh nghiệm, thường là không có nhận thức về khái niệm hệ quản trị cơ sở dữ liệu là gì. Họ truy nhập vào cơ sở dữ liệu thông qua những chương trình ứng dụng được viết phục vụ mục đích của họ, những chương trình này khiến cho việc sử dụng hệ thống càng đơn giản càng tốt. Họ thường không biêt gì, không có một chút kiến thức nào về cơ sở dữ liệu hoặc hệ quản trị cơ sở dữ liệu. * Nhóm thứ hai là nhóm sử dụng có kinh nghiệm hơn, là những người quen thuộc với cấu trúc của cơ sở dữ liệu và các phương tiện thực hiện được các chức năng do hệ quản trị cơ sở dữ liệu cung cấp. Họ thường sử dụng một ngôn ngữ truy vấn ở mức cao kiểu như SQL để thực hiện các thao tác cần thiết hoặc thậm cho tự viết các chương trình ứng dụng nhằm mục đích chuyên dụng đó. |
| 1.2.3. Qui trình thiết kế cơ sở dữ liệu |
| Quá trình thiết kế một cơ sở dữ liệu có thể chia thành sáu bước cơ bản sau. Các mô hình ngữ nghĩa dữ liệu liên quan nhiều tới ba bước đầu tiên.   1. Phân tích yêu cầu của bài toán: Đây là bước đầu tiên trong quá trình thiết kế một ứng dụng cơ sở dữ liệu để có thể hiểu được dữ liệu nào cần được lưu trữ trong cơ sở dữ liệu, ứng dụng nào cần được xây dựng để sử dụng chúng, và các thao tác dữ liệu nào cần được thực hiện thường xuyên và các yêu cầu về tốc độ thực hiện của hệ thống. Đây thường là tiến trình không chính thức liên quan tới những trao đổi với các nhóm người dùng và nghiên cứu môi trường hiện tại. Tiến hành tìm hiểu các ứng dụng hiện có cần được thay thế hoặc bổ trợ cho hệ thống cơ sở dữ liệu. 2. Thiết kế cơ sở dữ liệu mức khái niệm: Thông tin được thu thập trong bước phân tích yêu cầu được dùng để phát triển một bản mô tả mức tổng quát các dữ liệu cần được lưu trữ trong cơ sở dữ liệu, cùng với các ràng buộc cần thiết trên những dữ liệu này. 3. Thiết kế cơ sở dữ liệu ở mức logic: Một hệ quản trị cơ sở dữ liệu phải được lựa chọn để cài đặt một cơ sở dữ liệu và để chuyển đôiỉ bản thiết kế cơ sở dữ liệu mức khái niệm sang lược đồ cơ sở dữ liệu với mô hình dữ liệu của hệ quản trị cơ sở dữ liệu đã được lựa chọn. 4. Cải thiện các lược đồ: Trong bước nay các lược đồ được phát triển ở bước 3 được phân tích để phát hiện ra các vấn đề tiềm ẩn. Tại bước này, các lược đồ sẽ được chuẩn hóa. Việc chuẩn hóa một cơ sở dữ liệu được dựa trên một lý thuyết toán học rất mạnh và đẹp đẽ. Chúng ta sẽ bàn luận về việc chuẩn hóa này vào những buổi giảng sau. 5. Thiết kế cơ sở dữ liệu vật lý: Tại giai đoạn này, khối lượng công việc tiềm ẩn và các cách thức truy nhập được mô phỏng để xác định những điểm yếu tiềm ẩn trong cơ sở dữ liệu khái niệm. Quá trình này thường là nguyên nhân tạo ra các tệp chỉ mục hoặc/và các quan hệ phân cụm. Trong các tình huống sống còn, toàn bộ mô hình khái niệm sẽ cần được cấu trúc lại. 6. Thiết kế an toàn bảo mật cho hệ thống: Các nhóm người dùng khác nhau được xác định và các vai trò khác nhau của họ được phân tích sao cho cách thức truy nhập tới dữ liệu có thể xác định được. 7. Thông thường quá trình phát triển sẽ có bước thứ bảy hay là  bước cuối cùng, được gọi là giai đoạn chuốt lại hệ thống. Trong giai đoạn này, hệ thống cơ sở dữ liệu sẽ được thực hiện (mặc dù có thể nó chỉ được chạy trên chế độ mô phỏng) và sẽ được trau chuốt, cải thiện để đáp ứng được nhu cầu thực thi trong môi trường mong đợi. Hình vẽ dưới đây sẽ tóm tắt các bước chính trong quá trình thiết kế cơ sở dữ liệu. |
| 1.3. Mô hình trừu tượng 3 lớp |
| 1.3.1. Các mức trừu tượng |
| Để hiểu được quá trình thiết kế một hệ thống cơ sở dữ liệu, chúng ta cần xem xét một khái niệm mô hình trừu tượng 3 lớp trong một hệ thống cơ sở dữ liệu. Mô hình ba lớp này được thể hiện trong hình vẽ dưới đây   |  | | --- | |  | |  |   **Mức bên ngoài** là mức khung nhìn của người dùng đối với cơ sở dữ liệu, mô tả các cách nhìn khác nhau, những yêu cầu khác nhau của người dùng đối với hệ thống dữ liệu. Ở mức này, mỗi khung nhìn là một cách thể hiện một phần của cơ sở dữ liệu tương ứng với người sử dụng. Mỗi người sử dụng cần hệ thống theo một cách thức khác nhau và có một cách khác nhau, quen thuộc với mỗi người để biểu diễn thế giới thực được mô tả trong hệ thống. Khung nhìn từ bên ngoài bao gồm các thực thể, các thuộc tính của chúng và mối quan hệ giữa chúng trong thế giới thực mà người sử dụng đang quan tâm. Các thực thể khác, thuộc tính khác và các mối quan hệ khác có thể tồn tại nhưng không liên quan tới thế giới nhỏ đang cần mô tả đó thì người sử dụng sẽ không cần nhận thức sự tồn tại đó và không cần đưa chúng vào cơ sở dữ liệu.  Thường thì các khung nhìn khác nhau sẽ có cách biểu diễn khác nhau cho cùng một dữ liệu. Ví dụ một khung nhìn thể dữ liệu ngày dưới dạng (tháng, ngày, năm) trong khi một khung nhìn khác thể hiện ngày dưới dạng (ngày, tháng, năm). Một vài khung nhìn có thể bao gồm các dữ liệu phái sinh hoặc dữ liệu tính toán từ những dữ liệu khác. Những dữ liệu phái sinh này thực tế không được lưu trữ trong cơ sở dữ liệu mà chúng được tạo ra khi cần thiết. Ví dụ một khung nhìn có thể cần tuổi của một người. Tuy nhiên dữ liệu về tuổi không cần thiết phải lưu trữ trong cơ sở dữ liệu vì nó được cập nhật hàng. Thay vào đó nó có thể được tính toán từ dữ liệu được lưu trữ trong cơ sở dữ liệu biểu diễn ngày sinh của người đó và ngày tháng của hệ thống. **Mức khái niệm** là khung nhìn của những người thiết kế cơ sở dữ liệu. Mức này sẽ mô tả những dữ liệu cần thiết nào sẽ được lưu trữ trong đó và mối quan hệ giữa chúng.  Ở mức này, cấu trúc logic của toàn bộ cơ sở dữ liệu, được nhìn thấy bởi người quản trị cơ sở dữ liệu, được xác định. Nó thể một khung nhìn đầy đủ về các yêu cầu dữ liệu của một tổ chức mà không phụ thuộc vào bất kỳ một cách thức lưu trữ nào. Nói một cách khác, mức khái niệm không quan tâm tới việ lưu trữ vật lý của dữ liệu trong hệ thống, chỉ quan tâm tới việc xác định dữ liệu cần lưu trữ. Mức khái niệm hỗ trợ cho mỗi khung nhìn từ bên ngoài trong đó bất kỳ dữ liệu sẵn có nào tới người sử dụng cũng cần được lưu lại hoặc có khả năng sinh ra dữ liệu khác ở mức khái niệm. Mức này sẽ không chứa đựng những thông tin liên quan tới sự phụ thuộc vào cách lưu trữ vật lý trong hệ thống. Ví dụ, một thực thể có thể được xác định bằng cách thể hiện nó như một số nguyên tại mức này nhưng số byte mà nó chiếm giữ không cần được xác định cụ thể tại mức này. **Mức trừu tượng bên trong** thể hiện sự biểu diễn vật lý của cơ sở dữ liệu trong máy tính. Mức này sẽ mô tả cách thức dữ liệu được lưu trữ trong cơ sở dữ liệu. Mức trừu tượng này cũng sẽ mô tả sự cài đặt vật lý cần thiết để đạt được công suất chạy tối ưu và việc sử dụng không gian lưu trữ tốt nhất. Nó sẽ bao gồm cả các cấu trúc dữ liệu và việc tổ chức các tệp dữ liệu được sử dụng để lưu dữ liệu trong các thiết bị lưu trữ. Nó cung cấp giao diện với các phương thức truy nhập của hệ điều hành ( Đó là các kỹ thuật quản lý tệp liên quan tới việc lưu trữ và lấy ra các bản ghi dữ liệu) để đưa dữ liệu vào các thiết bị lưu trữ, xây dựng các tệp chỉ mục, lấy dữ liệu ra và các công việc khác. **Mức trừu tượng vật lý** là mức nằm bên dưới mức bên trong, mức này được quản lý bởi hệ điều hành dưới sự chỉ dẫn của hệ quản trị cơ sở dữ liệu. Các chức năng của hệ quản trị cơ sở dữ liệu và hệ điều hành ở mức vật lý này không có ranh giới rõ ràng và sẽ thay đổi từ hệ thống này sang hệ thống khác. Một vài hệ quản trị cơ sở dữ liệu tận dụng các ưu điểm của các phương pháp truy nhập của hệ điều hành, trong khi một số hệ quản trị khác sẽ chỉ sử dụng các phương pháp cơ bản và tự tạo ra những kiểu tổ chức tệp của riêng chúng.  Mức trừu tượng vật lý dưới hệ quản trị cơ sở dữ liệu bao gồm các thành phần mà chỉ được biết đến bởi hệ điều hành, ví dụ như việc tạo chuỗi thực hiện các công việc được diễn ra thế nào và liệu các trường của một bản ghi trong cơ sở dữ liệu có được lưu trữ thành các byte liền nhau trên đĩa không. |
| 1.3.2. Khái niệm lược đồ, ánh xạ và thể hiện của cơ sở dữ liệu |
| Lược đồ cơ sở dữ liệu là một mô tả tổng quát toàn bộ cơ sở dữ liệu. Có ba loại lược đồ khác nhau và chúng được định nghĩa dựa trên các mức độ trừu tượng của kiến trúc trừu tượng ba lớp:   * Ở mức độ cao nhất, có rất nhiều lược đồ mức ngoài. Mỗi lược đồ này được gọi là một lược đồ con dữ liệu, liên quan tới các khung nhìn khác nhau của dữ liệu * Ở mức khái niệm có một lược đồ khái niệm, sẽ mô tả tất cả các thực thể, thuộc tính và các mối quan hệ cùng với những ràng buộc toàn vẹn của chúng. * Mức trừu tượng thấp nhất sẽ có một lược đồ trong, sẽ là một mô tả hoàn thiện ở mô hình bên trong, bao gồm định nghĩa các bản ghi lưu trữ, các phương thức biểu diễn, v.v…   Hệ quản trị cơ sở dữ liệu chịu trách nhiệm ánh xạ giữa ba loại lược đồ này. Nó phải kiểm tra tính đồng nhất giữa ba loại đó; nói một cách khác, hệ quản trị cơ sở dữ liệu phải kiểm tra xem mỗi lược đồ ngoài có thể suy ra từ lược đồ khái niệm và nó phải sử dụng các thông tin trong lược đồ khái niệm để ánh xạ từ lược đồ ngoài vào các thành phần trong lược đồ trong. Lược đồ mức khái niệm có sự gắn kết với lược đồ bên trong thông qua một chuyển đổi được gọi là phép ánh xạ mức khái niệm/mức trong. Việc ánh xạ này cho phép hệ quản trị cơ sở dữ liệu tìm thấy bản ghi thực tế hoặc sự kết hợp của các bản ghi trong bộ lưu trữ vật lý đóng góp một bản ghi logic trong lược đồ khái niệm, cùng với các ràng buộc được gắn vào các phép toán liên quan tới bản ghi đó. Mỗi lược đồ bên ngoài có liên kết với lược đồ khái niệm với một phép chuyển đổi được gọi là phép ánh xạ bên ngoài/khái niệm. Phép chuyển đổi này cho phép hệ quản trị cơ sở dữ liệu ánh xạ các tên trong khung nhìn của người dùng lên các phần liên quan với lược đồ khái niệm. Ví dụ về việc chuyển đổi lược đồ giữa các mức trừu tượng đươc thể hiện như sau:   |  | | --- | |  | |  |   Sự độc lập dữ liệu   |  | | --- | |  | |  |   Một trong những mục tiêu chính của việc đưa ra kiến trúc trừu tượng 3 lớp là để cung cấp sự độc lập về dữ liệu, có nghĩa là lược đồ ở mức cao sẽ không bị ảnh hưởng tới sự  thay đổi của lược đồ ở mức thấp hơn. Tương ứng với ba mức độ trừu tượng sẽ có hai loại độc lập dữ liệu: độc lập dữ liệu mức logic (mức khái niệm) và mức vật lý. Độc lập dữ liệu mức khái niệm thể hiện ở sự không bị ảnh hưởng của các lược đồ mức ngoài đối với các thay đổi của lược đồ khái niệm. Điều này có nghĩa là nếu các thực thể, hay thuộc tính ở lược đồ mức khái niệm thay đổi, thì cũng không làm thay đổi nội dung các khung nhìn của người sử dụng. Độc lập dữ liệu mức vật lý thể hiện ở sự không bị ảnh hưởng của lược đồ khái niệm đối với những thay đổi của lược đồ mức vật lý. |
| 1.4. Các ngôn ngữ cơ sở dữ liệu |
| Một ngôn ngữ con dữ liệu bao gồm hai phần: một ngôn ngữ định nghĩa dữ liệu (Data Definition Language) và một ngôn ngữ thao tác dữ liệu (Data Manipulation Language). Ngôn ngữ định nghĩa dữ liệu được dùng để xác định lược đồ cơ sở dữ liệu chẳng hạn dùng để định nghĩa một quan hệ (một thực thể), các thuộc tính của thực thể đó, các kiểu dữ liệu của mỗi thuộc tính. Ngôn ngữ thao tác dữ liệu được dùng để đọc và cập nhật dữ liệu. Các ngôn ngữ này được gọi là ngôn ngữ con dữ liệu bởi vì chúng không bao gồm các cấu trúc lập trình cần thiết cho việc tính toán như cấu trúc điều kiện hoặc câu lệnh lặp, những cấu trúc được cung cấp bởi các ngôn ngữ lập trình bậc cao. Hầu hết các hệ quản trị cơ sở dữ liệu đều có một môi trường cho phép nhúng các ngôn ngữ con vào ngôn ngữ lập trình mức cao như COBOL, Pascal, C, C++, Java hoặc Visual Basic, những loại ngôn ngữ lập trình bậc cao này được gọi là ngôn ngữ chủ. Hầu hết các ngôn ngữ con cũng cung cấp một phiên bản tương tác hay không nhúng vào ngôn ngữ khác mà được đưa vào trực tiếp từ một thiết bị đầu cuối. **Ngôn ngữ định nghĩa dữ liệu** là một ngôn ngữ cho phép một quản trị cơ sở dữ liệu DBA hoặc người dùng mô tả và đặt tên các thực thể, các thuộc tính và các mối quan hệ cần thiết cho ứng dụng, cùng với các ràng buộc về bảo mật và toàn vẹn liên quan. Kết quả của việc thực thi hay biên dịch một câu lệnh định nghĩa dữ liệu là một tập các bảng được thu thập lưu trữ trong các tệp đặc biệt được gọi tên là bảng liệt kê các thành phần của hệ thống (system catalog). Tệp này cũng thường được gọi với cái tên là từ điển dữ liệu hay thư mục dữ liệu. **Ngôn ngữ thao tác dữ liệu** là một ngôn ngữ cung cấp một tập các thao tác hỗ trợ cho các phép toán thao tác dữ liệu trên các dữ liệu được lưu trữ trong cơ sở dữ liệu. Các thao tác của một ngôn ngữ thao tác dữ liệu thường bao gồm:   * Chèn thêm một dữ liệu mới vào cơ sở dữ liệu * Thay đổi dữ liệu được lưu trữ trong cơ sở dữ liệu * Lấy dữ liệu từ cơ sở dữ liệu ra * Xóa dữ liệu từ cơ sở dữ liệu   **Ngôn ngữ truy vấn** là một phần quan trọng của ngôn ngữ thao tác dữ liệu liên quan tới việc lấy dữ liệu ra từ cơ sở dữ liệu. Các ngôn ngữ thao tác dữ liệu được phân biệt bởi các cấu trúc lấy dữ liệu bên trong của nó, và được phân chia làm hai loại chính: có thủ tục và không thủ tục. Loại ngôn ngữ thao tác dữ liệu có thủ tục là các ngôn ngữ mà trong đó người dùng có thông báo với hệ thống những dữ liệu nào cần thiết và cách thức chính xác để lấy dữ liệu ra. Loại ngôn ngữ thao tác dữ liệu không có thủ tục là các ngôn ngữ trong đó người dùng chỉ thông báo cho hệ thống dữ liệu nào được yêu cầu và để hệ thống tự xác định cách thức lấy dữ liệu đó ra cho người sử dụng. Thông thường thì các ngôn ngữ thao tác dữ liệu có thủ tục sẽ được nhúng vào các ngôn ngữ lập trình mức cao. Các ngôn ngữ thao tác dữ liệu có thủ tục có xu hướng tập trung vào từng bản ghi đơn trong khi loại ngôn ngữ không thủ tục có xu hướng thực hiện trên một tập các bản ghi. **Các ngôn ngữ thế hệ thứ tư**: Chúng ta không có khái niệm về cái gì góp phần vào hình thành một ngôn ngữ thế hệ thứ tư. Cảm giác chung nó là một ngôn ngữ lập trình rất nhanh, những yêu cầu được thực hiện với hàng trăm dòng lệnh trong ngôn ngữ thế hệ thứ ba sẽ được thể hiện chỉ trong một vài dòng mã nguồn của ngôn ngữ thế hệ thứ tư. Ngôn ngữ thế hệ thứ ba là thuộc loại thủ tục, trong khi ngôn ngữ thế hệ thứ tư là loại không có thủ tục (đặc tính của không có thủ tục được trình bày rõ ở trên). Loại ngôn ngữ thế hệ thứ tư bao gồm các ngôn ngữ làm việc trên bảng tính và trên cơ sở dữ liệu. SQL và QBE (hai loại ngôn ngữ con dữ liệu liên quan tới bài giảng này) là hai ví dụ của ngôn ngữ thế hệ thứ tư. |