Tạp chí Tin học và Điều khiển học, T. 17, S. 2 (2001), 75–81

VỀ MỘT CÔNG CỤ KHAI THÁC CƠ SỞ DỮ LIỆU ORACLE - DISCOVERER

TRẦN THỊ PHIẾN

Abstract. Oracle - Discoverer is a tool that supports data mining from the large databases and creates the dynamic reports with writing SQL statements. In this paper we give some principal conceptions and the way of exploiting database by Oracle - Discoverer.

Tóm tắt. Oracle - Discoverer là một công cụ hỗ trợ khai thác thông tin từ các CSDL lớn tạo ra các báo cáo động với sự trợ giúp của các câu lệnh SQL. Bài báo này giới thiệu một số khái niệm cơ bản và cách khai thác dữ liệu bằng Oracle-Discoverer.

1. MỞ ĐẦU

Hiện nay, nhu cầu truy cập thông tin ngày càng tăng, dung lượng và mức độ phức tạp của dữ liệu cũng như số lượng ứng dụng phát triển một cách nhanh chóng. Nếu như trước đây người sử dụng chỉ cần các chương trình đơn giản truy cập các tệp dữ liệu thì bây giờ cần phải có các công cụ mạnh hỗ trợ trong việc khai thác thông tin phục vụ công việc chuyên môn của mình.

Oracle - Discoverer là một trong những công cụ đó. Đối với người sử dụng công cụ hỗ trợ trong việc tìm kiếm, phân tích dữ liệu cần thiết từ kho dữ liệu không chỉ theo nhiều chiều mà còn theo chiều sâu nhằm đưa ra những quyết định đúng đắn trong công tác quản lý. Còn đối với người quản trị hệ thống Discoverer cho phép họ tạo ra những tệp con dữ liệu thuộc các lĩnh vực khác nhau cần thiết cho các chuyên viên để hỗ trợ việc tạo ra quyết định.

Bài này giới thiệu một số khái niệm cơ bản và cách khai thác dữ liệu bằng Oracle - Discoverer.

2. GIỚI THIỆU CHUNG VỀ CÔNG CỤ DISCOVERER

Oracle - Discoverer (OD) của hãng Oracle hỗ trợ khai thác thông tin đa chiều, tạo ra các báo cáo động với sự trợ giúp các câu lệnh SQL. Một trong những kiểu phân tích phổ biến nhất là phân tích dữ liệu đa chiều và phân tích dữ liệu theo chiều sâu. Khai thác bằng OD có tính mở bởi người quản trị hệ thống hoàn toàn chủ động trong việc thêm, bớt các vùng tác nghiệp, tạo thêm những phân rã, những lớp mục mới, các trang tính mới... cho phù hợp với nhu cầu khai thác của các đối tượng sử dụng.

Trong quá trình nghiên cứu, nhóm phát triển ứng dụng đã khai thác triệt để công cụ và áp dụng thử nghiệm vào việc khai thác dữ liệu của CSDL chủ đề thu chi ngân sách của Bộ Tài chính được xây dựng theo công nghệ Kho dữ liệu (Data Warehousing) bước đầu đạt kết quả tốt. Các kho dữ liệu từ các hệ quản trị CSDL khác như Visual Fox, Foxpro... theo các khuôn dạng khác nhau (DBF, TXT, XLS) được đưa vào CSDL thông qua bộ công cụ trợ giúp việc chuyển đổi dữ liệu để phục vụ công việc khai thác dữ liệu không chỉ từ một CSDL hiện hữu mà có thể từ nhiều nguồn dữ liệu khác nữa. Để có thể tiến hành thử nghiệm được, người sử dụng cần cài đặt phần mềm công cụ này trên máy trạm trong môi trường của Oracle. Dưới đây là một số khái niệm cơ bản của Oracle-Discoverer.

Vùng tác nghiệp (Business Area - BA) gồm một số bảng của một CSDL nào đó, các bảng này có quan hệ với nhau để cùng thực hiện một nhiệm vụ theo yêu cầu của công việc. Trong vùng tác nghiệp ta có thể tạo ra các điều kiện lọc dữ liệu, các liên kết giữa các bảng, tạo ra những phân cấp dữ liệu theo các chiều như chiều thời gian, địa bàn...

Vùng tác nghiệp do người quản trị CSDL tạo ra và phân quyền sử dụng cho người dùng cuối (End User - EU).

Trang tính (Worksheet) chứa các khuôn dạng báo cáo để hiển thị dữ liệu theo ý muốn của người dùng cuối với những thao tác đơn giản. Người dùng cuối thường không phải là những chuyên gia tin học, họ chỉ cần biết thông tin để đưa ra quyết định, phân tích, báo cáo... Những người dùng cuối này sẽ làm việc trên các vùng tác nghiệp mà những chuyên gia tin học đã tạo ra cho họ. OD giúp họ tạo ra được những trang tính chứa những khuôn dạng báo cáo để hiển thị thông tin theo ý muốn. Công việc này thực hiện bằng những thao tác đơn giản, không cần phải gõ lệnh mà chỉ cần kích chuột. Ngoài ra, OD còn cung cấp môi trường tạo lập các bảng tính cho người sử dụng.

Tầng người dùng cuối (End User Layer - EUL) làm nhiệm vụ trung gian giữa CSDL và người dùng cuối. EUL có nhiệm vụ tự động phát sinh ra câu lệnh để lấy dữ liệu chuyển cho người dùng cuối. Mỗi người dùng sẽ phải tạo ra một tầng EUL trung gian khi làm việc. Trong số đó, chỉ có một người dùng cuối tạo ra tầng EUL công cộng (public), còn những người dùng khác chỉ có thể tạo ra các EUL riêng (private). Có thể nói EUL là hạt nhân của bộ công cụ Discoverer.

Để có thể sử dụng được OD trên một CSDL nào đó, trước tiên phải chạy bộ công cụ "Install End User Layer Tables 3.0" để tạo ra một tầng người sử dụng cuối công cộng, sau đó tạo một số tầng người sử dụng riêng tùy theo mức độ cần thiết.

Phần thao tác của người sử dụng được thiết kế cho những chuyên gia về nghiệp vụ, không có kiến thức về lập trình trên máy tính cũng như kiến thức về CSDL. Oracle Discoverer User Edition là một công cụ truy nhập dữ liệu rất dễ sử dụng. Nó cung cấp những truy nhập logic và trực giác tới thông tin từ CSDL quan hệ cho những báo cáo, phân tích và những truy vấn đặc biệt (những truy vấn này không được xác định từ trước như đưa ra báo cáo theo những biểu mẫu sẵn có mà đây là những truy vấn trực tiếp từ người sử dụng phát sinh ngay tại chỗ nhằm mục đích hỗ trợ quyết định).

Phần quản trị tạo dựng và duy trì EUL. Thiết kế của nó quyết định cách thức người sử dụng truy nhập và xem dữ liệu như thế nào.

Tầng trung gian với người sử dụng cuối: Các CSDL quan hệ thường rất phức tạp và chứa hàng trăm bảng. Thêm nữa thiết kế CSDL thường xuyên thay đổi để đáp ứng được sự tăng trưởng của khối lượng công việc. EUL tách người sử dụng ra khỏi sự phức tạp của CSDL và sự thay đổi cố định đó. Nó cung cấp một khung nhìn trực quan, hướng tới công việc của CSDL rất phù hợp cho mỗi người sử dụng hoặc mỗi nhóm người sử dụng. Như vậy EUL hướng người sử dụng quan tâm tới vấn đề nghiệp vụ phục vụ cho công việc của họ thay vì quan tâm tới vấn đề truy nhập dữ liệu.

Từ một truy vấn của người sử dụng, EUL tạo ra những câu lệnh SQL trên máy client và kết nối với CSDL thông qua SQL*Net. Khi một người sử dụng chọn các folder (bảng dữ liệu) và các item (các cột trong bảng), EUL tạo ra những câu lệnh SQL tương ứng để xác định sự lựa chọn từ bảng, khung nhìn hoặc cột nào. Khi người sử dụng thực hiện truy vấn, EUL sinh ra những câu lệnh SQL và gửi chúng tới CSDL, sau đó CSDL sẽ gửi trả lại kết quả tới giao diện với người sử dụng cuối.

Vì vậy, người sử dụng cuối không cần phải hiểu bất kì một câu lệnh SQL nào dùng để truy nhập, phân tích và lấy dữ liệu ra. Tất cả những công việc đó đều do EUL thực hiện và khiến cho chúng trở nên trong suốt đối với người sử dụng.

Metadata mô tả cấu trúc dữ liệu, nội dung, khóa, chỉ mục, phương pháp xử lý, phương pháp tổ chức dữ liệu...

Metadata được chia thành 3 loại: công việc, kĩ thuật và tác nghiệp.

Metadata công việc chứa đựng những thông tin giúp người sử dụng dễ dàng hiểu được khung cảnh của thông tin được lưu trữ trong kho:

- Các vùng chủ thể và các loại đối tượng thông tin bao gồm các câu truy vấn, các báo cáo, các hình ảnh...

- Các thông tin khác để hỗ trợ cho tất cả các thành phần cấu thành kho dữ liệu. Chẳng hạn như các thông tin liên quan tới các hệ thống phân phối thông tin bao gồm các thông tin về lịch làm việc, những chi tiết về nơi phân phối, các truy vấn, báo cáo và các phân tích được xác định trước.

- Các thông tin tác nghiệp của kho dữ liệu như lịch sử của dữ liệu (các snapshot, các version), quyền sở hữu, theo dõi sổ sách, sử dụng dữ liệu.

- Miêu tả các thuộc tính kho dữ liệu bằng cách xác định tên của công việc, các định nghĩa, các bảng mô tả và các bí danh.

Metadata kĩ thuật chứa đựng những thông tin về dữ liệu trong kho của những người thiết kế và quản trị khi tiến hành công việc phát triển và quản lý:

- Thông tin về các nguồn dữ liệu từ các hệ thống tác nghiệp và những hệ thống bên ngoài môi trường kho dữ liệu về vị trí, tên các file, kiểu file, tên các trường và các đặc tính, bí danh, thông tin về phiên bản, những mối quan hệ, độ lớn, tính dễ biến động, người chủ dữ liệu và những người sử dụng có quyền truy nhập.

- Những mô tả về sự chuyển đổi ví dụ như cách thức ánh xạ từ CSDL tác nghiệp vào kho dữ liệu và các thuật toán được sử dụng để biến đổi và cải thiện hay chuyển đổi dữ liệu.

- Những định nghĩa cấu trúc dữ liệu và đối tượng trong môi trường kho dữ liệu cho dữ liệu đích.

- Những luật dùng để làm sạch và trích lọc dữ liệu.

- Quyền truy nhập, lịch sử về backup, lưu trữ, phân phối, thu nhập dữ liệu, v.v..

Metadata tác nghiệp (Operational Metadata - OM)

- Trợ giúp trong việc duy trì và triển khai kho dữ liệu.

- Mô tả thông tin chứa đựng trong các bảng đích.

- Mô tả cốt lõi, khả năng tạo CSDL đích (tạo ra bảng và thông tin dưới dạng liệt kê), thông tin được lưu trữ hay trực tuyến, ngày refresh, số lượng các bản ghi, lịch thực hiện các công việc và những người sử dụng có khả năng truy nhập vào dữ liệu.

- Cung cấp các thông tin về dữ liệu, chẳng hạn thời gian dữ liệu được tập hợp lại trong bảng đích, thời điểm các công việc được thực hiện theo kế hoạch và thực sự được thực hiện, bảng đích được tải vào lần cuối cùng, số lượng bản ghi được tải vào, truy vấn chung được thực hiện trên một bảng, bảng có đòi hỏi một chỉ mục (index) nào khác không.

Metadata cung cấp khả năng giao tiếp với người sử dụng cuối cùng về những thông tin bên trong kho và cách thức chúng được truy nhập để giúp cho họ có thể hiểu được nội dung và tìm thấy được dữ liệu cần thiết.

Việc lưu trữ, quản lý, và phân loại metadata được thực hiện qua một kho chứa metadata và các phần mềm kèm theo. Các kho được phân loại bằng cách sử dụng một sơ đồ phân loại được gọi là *mô hình thông tin* (*information model*). Mô hình này chứa một danh sách các loại siêu dữ liệu và sự liên quan giữa chúng. Kho này là một thiết bị quản lý siêu dữ liệu với mục đích chung và rất linh hoạt. Các siêu dữ liệu được lưu trữ và quản lý bởi kho siêu dữ liệu. Phần mềm quản lý kho siêu dữ liệu có thể được sử dụng để ánh xạ dữ liệu nguồn tới CSDL đích, tạo ra mã của việc chuyển đổi dữ liệu, tích hợp, chuyển đổi và kiểm soát sự dịch chuyển dữ liệu vào trong kho. Các phần mềm này chạy trên máy trạm và cho phép người sử dụng biết được dữ liệu được chuyển đổi như thế nào, ví dụ như ánh xạ, biến đổi hay tổng hợp. Metadata cung cấp các con trỏ hướng hỗ trợ quyết định trỏ tới kho và cung cấp một liên kết logic giữa kho dữ liệu và ứng dụng hỗ trợ quyết định. Một kho dữ liệu được thiết kế để đảm bảo có một cơ chế sản sinh và duy trì kho siêu dữ liệu và tất cả các đường dẫn truy nhập vào kho dữ liệu đều thông qua metadata.

3. QUẢN TRỊ HỆ THỐNG ĐỐI VỚI DISCOVERER

Một người quản trị hệ thống cần nắm được cách sử dụng CSDL để trợ giúp cho việc ra quyết định của tổ chức. Thêm nữa, cần phải hiểu được dữ liệu trong CSDL, chúng được định vị ở đâu, được lưu trữ như thế nào và mối liên hệ giữa chúng (kể cả mối liên hệ với những dữ liệu khác). Về

phương diện nghiệp vụ cần phải nắm được dữ liệu mà những người sử dụng dùng chúng để trợ giúp việc ra quyết định: yêu cầu, kiểu phân tích dữ liệu cần thiết và kết quả cuối cùng biểu diễn dưới dạng nào cho dễ dàng nhận biết và hiểu được.

Do đó cần phải phỏng vấn những người sử dụng cuối cùng để tìm ra được các kiểu phân tích, các dữ liệu (nằm trong CSDL của tổ chức nhưng hạn chế trong một phạm vi nào đó) mà họ cần. Công việc của người quản trị hệ thống là đáp ứng được nhiều yêu cầu của người sử dụng, có thể tạo được những trang tính liên quan tới một hay nhiều vùng tác nghiệp.

Ngoài ra người quản trị hệ thống có trách nhiệm bảo đảm vấn đề bảo mật. Cần kiểm soát sự truy nhập của người sử dụng cuối tới các vùng tác nghiệp. Vùng tác nghiệp thiết lập ra một tầng bảo mật thứ hai (tầng thứ nhất là bảo mật của CSDL). Tất cả truy nhập chính tới các đối tượng của CSDL (như các bảng hay các view) đều được kiểm soát bởi người quản trị CSDL.

Công việc quản trị bao gồm:

- Xác định các đối tượng của CSDL có thể được gộp nhóm một trong một vùng tác nghiệp.
- Tạo ra tên có ý nghĩa và gợi nhớ cho vùng tác nghiệp, các folder và các item.
- Kiểm soát truy nhập của người sử dụng cuối (hay nhóm người sử dụng cuối) tới các vùng tác nghiệp.
- Phân quyền sử dụng các chức năng như quản trị, tạo các bảng tổng hợp trước, tạo các vùng tác nghiệp...
- Xác định các công thức và thuộc tính của người sử dụng.
- Tạo ra các điều kiện kết hợp và những folder ghép mới.
- Tạo ra những điều kiện và những phân rã các item dùng cho việc khai phá dữ liệu theo chiều sâu để trợ giúp phân tích của người sử dụng cuối.
- Tạo ra những bảng tổng hợp trước (bảng summary).

4. KHAI THÁC DỮ LIỆU BẰNG DISCOVERER

Ở đây, không đề cập tới việc phỏng vấn người sử dụng để tạo ra đầy đủ các vùng làm việc hiệu quả cho tất cả các đối tượng sử dụng (vì đây là công việc của người quản trị hệ thống của tổ chức đó) mà chỉ có ý định mô phỏng CSDL được thiết kế để tổng hợp và lưu trữ dữ liệu tốt và hiệu quả cho việc khai thác thông tin bằng công cụ phân tích dữ liệu, tạo truy vấn đặc biệt và báo cáo khai thác theo nhiều chiều, khoan sâu dữ liệu từ tổng hợp tới chi tiết như công cụ Discoverer.

1. Với công cụ Discoverer người sử dụng có thể khai thác dữ liệu theo các vùng tác nghiệp tương ứng tùy theo mức độ chi tiết có thể.
2. Cung cấp cách thể hiện số liệu của các truy vấn bằng các báo cáo trực quan dưới dạng biểu đồ theo nhiều dạng khác nhau. Điều này hỗ trợ rất nhiều cho những người cần đưa ra quyết định chiến lược trong việc tìm ra xu hướng phát triển theo một lĩnh vực nào đó mà CSDL cung cấp.

Trong khi thiết kế CSDL, việc xác định các bảng trung tâm (Fact Tables - FT) có quan hệ chặt chẽ đến việc lưu trữ trong kho dữ liệu và thực hiện việc khai thác sau này. Bởi vậy, nếu FT là một bảng quá lớn hoặc phức tạp sẽ ảnh hưởng trực tiếp đến tốc độ xử lý thông tin. Việc xác định các bảng FT cần dựa vào thông tin được lưu trữ theo các nguồn dữ liệu đồng thời cũng phải dựa trên những nhu cầu khai thác của người sử dụng cuối.

Như ta đã biết, có một kĩ thuật để cải thiện tốc độ truy vấn là tạo ra các bảng FT kết hợp trước các chiều cần khai thác chính. Nếu số liệu càng được tổng hợp và tính toán trước theo mục đích khai thác thì tốc độ truy vấn càng được cải thiện có nghĩa là cần tạo thêm nhiều bảng FT mà mỗi bảng này có kích thước nhỏ hơn đồng thời phải có một FT lưu trữ tất cả các chiều cùng khai thác một lúc. Tùy thuộc vào nhu cầu khai thác thông tin mà tổ chức mô hình dữ liệu khai thác, tạo ra các bảng tạm để lưu trữ số liệu tổng hợp trước theo một số chiều nào đó để cải thiện tốc độ truy vấn, khai phá dữ liệu theo những chiều đó.

Mô hình dữ liệu thường được sử dụng cho hệ thông tin tác nghiệp là mô hình dữ liệu quan hệ,

một mô hình dựa trên các nguyên lí toán học và logic vị từ. Việc định nghĩa sơ đồ dữ liệu thường dựa trên tối đa sự đồng thời và tối ưu những thao tác xóa, thay đổi, chèn thêm thông qua việc xác định các bảng quan hệ tương ứng với những yêu cầu tác nghiệp và nội dung lưu trữ được tối thiểu nhất cho việc truy nhập tới từng bản ghi riêng.

Giải pháp để xây dựng một CSDL đa chiều có hiệu quả là phải kết hợp từ trước tất cả các tổng con logic và các tổng theo tất cả các chiều. Sự kết hợp trước này đặc biệt có giá trị khi các chiều mang tính phân cấp, giúp cho người sử dụng thực hiện khả năng khoan sâu (drill_down) dữ liệu - từ một nhóm các sản phẩm xuống từng sản phẩm riêng rẽ, từ việc bán hàng theo từng năm xuống theo tuần.

Sự phân cấp về kích thước, quản lý dữ liệu thưa hơn và sự kết hợp trước là quan trọng vì chúng làm giảm đáng kể kích cỡ CSDL và những yêu cầu tính toán các giá trị. Một thiết kế như vậy loại bỏ việc phải kết hợp nhiều bảng và cung cấp sự truy nhập trực tiếp và nhanh tới các câu trả lời vì vậy tăng tốc độ đáng kể trong việc thực hiện các truy vấn đa chiều.

Một số loại sơ đồ thông dụng được sử dụng trong thiết kế CSDL đa chiều, bao gồm:

Sơ đồ hình sao

Trong sơ đồ hình sao dữ liệu được xác định và phân loại 2 kiểu: sự kiện (bảng fact) và phạm vi (các bảng dimension). Các sự kiện là các đại lượng số của công việc. Các phạm vi là các bộ lọc hoặc các ràng buộc của những sự kiện này.

Vì bảng fact được tổng hợp từ trước và được kết hợp theo nhiều chiều nên xu hướng có rất nhiều hàng và tăng trưởng một cách nhanh chóng trong khi đó các bảng dimension không có nhiều hàng và sự tăng trưởng là tương đối tĩnh. Bảng fact có thể bao gồm hàng nghìn hàng. Bảng dimension (bảng theo chiều) chứa đựng các thuộc tính có thể được sử dụng như các tiêu chí tìm kiếm và thường có kích thước nhỏ hơn nhiều, rất quen thuộc với người sử dụng từ trước. Khóa của nó không là khóa ghép như bảng fact. Nếu một bảng dimension bắt đầu có sự tương đồng với các bảng fact thì có thể nó cần được chia ra thành các bảng dimension. Nếu một bảng dimension được chia ra thành dimension chính và dimension thứ 2 thì cấu trúc thu được của kết quả được coi là một snowflake (sơ đồ bông tuyết) hoặc một cấu trúc sao mở rộng.

Có nhiều loại sơ đồ hình sao từ đơn giản đến phức tạp. Một sơ đồ hình sao đơn giản chỉ gồm một bảng fact và một vài bảng theo chiều. Một sơ đồ hình sao phức tạp bao gồm hàng trăm bảng fact và bảng theo chiều.

Một vài kĩ thuật để cải thiện công suất của các truy vấn trong sơ đồ hình sao bao gồm:

- Xác định sự kết hợp các bảng fact đang tồn tại hay tạo ra một sự tập hợp mới dữ liệu từ các bảng fact để tạo ra một họ các bảng fact.
- Tập hợp dữ liệu là quá trình tích lũy dữ liệu của các bảng fact theo những thuộc tính được xác định trước (chính là việc tính tổng các số liệu của bảng fact để lưu trữ trong các bảng fact mới mà các bảng này chỉ bao gồm những thuộc tính xác định trước đó) để nhằm mục đích phục vụ yêu cầu người khai thác dữ liệu.
- Phân chia bảng fact đến mức mà hầu hết các truy vấn chỉ truy nhập tới phần đó.
- Tạo ra các bảng fact riêng rẽ.
- Tạo ra những tệp chỉ số đơn duy nhất hoặc các kĩ thuật khác để cải thiện năng suất kết hợp.

Lưu ý kể cả bảng fact và các bảng theo chiều đều không bắt buộc ở dạng chuẩn như đối với phương pháp thiết kế truyền thống tức là trong CSDL có dư thừa dữ liệu. Với loại sơ đồ này cho phép lưu trữ dư thừa dữ liệu đổi lại khả năng truy nhập nhanh hơn phù hợp với những câu hỏi phân tích nhiều chiều, phức tạp.

Sơ đồ bông tuyết

Sơ đồ bông tuyết là một sự mở rộng của sơ đồ hình sao tại đó mỗi cánh sao không phải là một bảng dimension mà là nhiều bảng. Trong dạng sơ đồ này, mỗi bảng theo chiều của sơ đồ hình sao được chuẩn hóa hơn. Sơ đồ bông tuyết cải thiện năng suất truy vấn, tối thiểu không gian đĩa cần

thiết để lưu trữ dữ liệu và cải thiện năng suất nhờ việc chỉ phải kết hợp những bảng kích thước nhỏ hơn thay vì phải kết hợp những bảng có kích thước lớn lại không chuẩn hóa. Nó cũng làm tăng tính linh hoạt của các ứng dụng bởi sự chuẩn hóa và ít mang bản chất theo chiều hơn. Nó làm tăng số lượng các bảng và làm tăng tính phức tạp của một vài truy vấn cần có sự tham chiếu tới nhiều bảng.

Sơ đồ kết hợp giữa hai loại trên

Đó là một sự kết hợp giữa sơ đồ hình sao dựa trên bảng fact và những bảng theo chiều không chuẩn hóa theo các chuẩn 1, 2, 3 và sơ đồ bông tuyết trong đó tất cả các bảng dimension đều đã được chuẩn hóa, trong sơ đồ loại này chỉ những bảng dimension lớn là được chuẩn hóa còn những bảng khác chứa một khối lượng lớn các cột dữ liệu chưa được chuẩn hóa.

Nhiều sơ đồ hình sao được gọi là một họ các sơ đồ hình sao. Nó là một khái niệm của sơ đồ hình sao làm cho mô hình loại này không thể quản lý được.

Một vài CSDL và các công cụ truy vấn của người sử dụng cuối nhất là các công cụ xử lý phân tích trực tuyến (OLAP) đòi hỏi mô hình dữ liệu phải là sơ đồ hình sao bởi vì nó là một mô hình dữ liệu quan hệ nhưng lại được thiết kế để hỗ trợ những thuộc tính của mô hình dữ liệu đa chiều là điểm cốt lõi của OLAP. Các CSDL và công cụ này được điều chỉnh cho phù hợp thực hiện được các yêu cầu truy vấn đối với mô hình này.

Một trong những cách để tăng công suất thực hiện các truy vấn của RDBMS là sử dụng kĩ thuật đánh chỉ số mới cho phép truy nhập nhanh, trực tiếp tới dữ liệu. Mỗi lần dữ liệu được tải vào, tất cả dữ liệu được chuyển đổi thành các chuỗi bitmap, những chuỗi này sau đó được nén lại và được lưu trữ trên đĩa. Khác với việc đánh chỉ số thông thường, những chỉ số không chỉ tới dữ liệu được lưu trữ ở nơi khác mà tất cả dữ liệu được lưu trữ trong cấu trúc chỉ số này. Tệp chỉ số bitmap có thể trở nên cồng kềnh và thậm chí không phù hợp đối với dữ liệu có lực lượng lớn khi phạm vi giá trị của dữ liệu là lớn. Một giải pháp khác là sử dụng cấu trúc chỉ số B-tree (cây nhị phân). Tuy nhiên, phương pháp này có thể làm tăng kích thước bởi vì khi khối lượng dữ liệu và số lượng các chỉ số tăng thì chúng đòi hỏi thường xuyên được duy trì khi dữ liệu được thêm vào, được cập nhật hay được xóa đi khỏi CSDL. Như vậy ta thấy rằng chỉ số B-tree có thể cải thiện một cách đáng kể công suất truy vấn nếu kiểu câu hỏi truy vấn được biết trước và tệp chỉ số được xây dựng để phản ánh đường dẫn truy nhập đã được biết trước. Nhưng B-tree không hiệu quả đối với những câu hỏi truy vấn đặc biệt (có thể hiểu là những truy vấn không biết trước) điển hình của các ứng dụng kho dữ liệu.

Các công cụ truy vấn dữ liệu

Những công cụ truy vấn dữ liệu khiến cho sự phức tạp của ngôn ngữ SQL và của cấu trúc CSDL là trong suốt với người dùng bằng cách chèn thêm vào một metalayer giữa người sử dụng và CSDL. Metalayer là một phần mềm cung cấp những khung nhìn (view) hướng chủ đề của một CSDL và hỗ trợ việc tạo ra các câu lệnh SQL bằng cách chọn và nhấn chuột (point-and-click). Chúng cũng hỗ trợ thực hiện những truy vấn không đồng bộ và việc tích hợp với Web server. Hãng Oracle đã đưa ra một phần mềm thuộc loại này là Discoverer/2000. Những công cụ với kiến trúc nhiều tầng làm việc với cơ chế chung như sau:

- Kiến trúc một tầng: Phần mềm client và CSDL nằm trên cùng một máy vật lí.
- Kiến trúc hai tầng: Phần mềm client và CSDL nằm trên hai máy khác nhau.
- Kiến trúc ba tầng: Phần mềm client và CSDL nằm trên hai máy khác nhau. Tầng thứ ba thay đổi tùy theo từng nhà cung cấp phần mềm, được sử dụng cho một hoặc nhiều mục đích: mô tơ tính toán, quản lý nguồn, CSDL đa chiều.

5. KẾT LUẬN

Kỹ thuật khai thác dữ liệu theo chiều sâu là quá trình đào xới, xem xét dữ liệu dưới nhiều mức

độ nhằm tìm ra mối liên hệ giữa các thành phần dữ liệu và phát hiện ra những xu hướng, hình mẫu và những kinh nghiệm quá khứ tiềm ẩn trong kho dữ liệu, vì vậy nó rất phù hợp với mục đích phân tích dữ liệu hỗ trợ cho công việc điều hành và ra quyết định.

TÀI LIỆU THAM KHẢO

[1] Vidette Poe, *Building a Data Warehouse for Decision Support*, Prentice Hall PTR, 1997.
[2] Bary Devlin, *Data Warehouse from Architecture to Implementation*, Addition Wessley, 1997.
[3] Harjinder S. Gill and Prakash Rao, *The Official Client/Server Computing Guide to Data Warehousing*, Que Corporation, 1996.
[4] Oracle, *Discoverer Release 3.0*.

Nhận bài ngày 22 tháng 10 năm 2000
Nhận bài sau khi sửa ngày 20 tháng 4 năm 2001

Viện Công nghệ thông tin