BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGUYỄN VĂN DŨNG

TỐI ƯU HÓA TRUY VẤN TRÊN CƠ SỞ DỮ LIỆU PHÂN TÁN

Chuyên ngành : Khoa học máy tính
Mã số : 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2012

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TSKH. TRẦN QUỐC CHIẾN**

Phản biện 1 : **PGS.TS. PHAN HUY KHÁNH**

Phản biện 2 : **GS.TS. NGUYỄN THANH THỦY**

Luận văn được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày tháng năm 2012

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng;
- Trung tâm Học liệu, Đại học Đà Nẵng;

MỞ ĐẦU

1. Lý do chọn đề tài

Khi thực thi một truy vấn, có thể có nhiều phương án mà hệ thống cơ sở dữ liệu (CSDL) cho phép xử lý và sản sinh câu trả lời. Các phương án có kết quả cuối cùng là tương đương về kết quả tính toán nhưng khác nhau trong chi phí thực hiện, lựa chọn phương án nào để có tổng chi phí thực hiện là nhỏ nhất? Khi truy vấn cơ sở dữ trong quá trình học hay thử nghiệm với dữ liệu nhỏ thì không ai quan tâm nhiều đến vấn đề này, nhưng khi dữ liệu đã lên tới cỡ triệu bản ghi thì vấn đề thời gian trả ra kết quả truy vấn lại là vấn đề lớn và cần có giải pháp hiệu quả.

Với cơ sở dữ liệu ngày càng đồ sộ, lưu trữ phân tán và việc vấn tin là vấn đề thường xuyên, để tạo ra một hoạch định thực thi vấn tin nhằm hạ thấp tối đa chi phí thì việc tối ưu câu vấn tin là vấn đề mà ai cũng phải quan tâm. Từ đó có thể nhận thấy rằng vấn đề tối ưu hoá truy vấn phân tán là cấp thiết trong các hệ quản trị CSDL.

2. Mục đích nghiên cứu

Đề tài phân tích, tổng hợp, bình luận và trình bày một cách có hệ thống các nghiên cứu về cơ sở dữ liệu quan hệ, hệ tin học phân tán, cách thiết kế cơ sở dữ liệu phân tán, trên cơ sở các ứng dụng truy vấn để đưa ra phương pháp thiết kế, tối ưu và chọn lọc chiến lược thực thi truy vấn hiệu quả nhất.

3. Đối tượng và phạm vi nghiên cứu

Tối ưu hóa truy vấn CSDL phân tán có ý nghĩa to lớn trong việc cải thiện tốc độ truy xuất, tìm kiếm thông tin, có thể có nhiều phương án để đưa ra kết quả nhưng nghiên cứu phương án tốn ít chi phí hơn là vấn đề được nhiều người quan tâm.

Đối tượng được nghiên cứu là câu truy vấn SQL cho CSDL tập trung, và tối ưu hóa câu truy vấn đó để sinh ra các mảnh ở những vị trí khác nhau nhằm tối ưu hóa về chi phí thực hiện. Đối tượng nghiên cứu đó thuộc phạm vi nghiên cứu lý thuyết về tối ưu hóa, ứng dụng trong lĩnh vực giáo dục, đào tạo là chủ yếu.

4. Phương pháp nghiên cứu

Phương pháp chính là nghiên cứu tài liệu, nghiên về lý thuyết truy vấn, chi phí trong quá trình truy vấn, có xây dựng một ứng dụng mô phỏng yêu cầu của vấn đề cần nghiên cứu.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Việc tối ưu hóa truy vấn trên cơ sở dữ liệu phân tán sẽ giúp cho việc nghiên cứu, khai thác ứng dụng trên hệ thống phân tán, chủ yếu là qua môi trường mạng được thuận lợi và phát triển hơn.

6. Bố cục luận văn

Toàn bộ nội dung của luận văn được chia thành các chương như sau:

Chương 1. Cơ sở lý thuyết

Chương này sẽ trình bày các nội dung về lý thuyết về hệ quản trị cơ sở dữ liệu, mô hình cơ sở dữ liệu, ngôn ngữ đại số quan hệ và một số khái niệm trong hệ tin học phân tán.

Chương 2. Thiết kế và tối ưu hóa truy vấn phân tán

Chương này trình bày các nội dung về lý thuyết để thiết kế cơ sở dữ liệu phân tán như các mục tiêu, chiến lược và một số vấn đề khi thiết kế, phân mảnh, cấp phát cho các mảnh... Đồng thời trong chương này sẽ trình bày các nguyên tắc tối ưu hóa, mô hình và các thuật toán tối ưu hóa.

Chương 3. Xây dựng hệ thống quản lý nhân viên

Chương này xây dựng hệ thống quản lý nhân viên, từ đó đưa ra các ứng dụng truy vấn để xác định việc phân mảnh, cấp phát phân mảnh và thành lập các dữ liệu phân tán, quyền truy cập vào các cơ sở dữ liệu đó.

CHƯƠNG 1
CƠ SỞ LÝ THUYẾT

1.1. HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU

1.1.1. Cơ sở dữ liệu

1.1.2. Hệ quản trị CSDL

- Hệ quản trị CSDL là hệ thống phần mềm có chức năng tạo lập và quản trị CSDL như cập nhật, thêm, sửa, xóa, sắp xếp, tìm kiếm, thống kê và quản lý các truy cập của người sử dụng đến cơ sở dữ liệu.
- Hệ quản trị CSDL phân tán là hệ thống phần mềm, cho phép quản lý các hệ CSDL phân tán và làm cho việc phân tán trở nên vô hình đối với người sử dụng.

1.1.3. Mô hình dữ liệu quan hệ

1.1.4. Ngôn ngữ đại số quan hệ

1.1.4.1. Phép hợp (Union)

1.1.4.2. Phép giao (Intersection)

1.1.4.3. Phép hiệu (Minus)

1.1.4.4. Phép chiếu (Projection)

Cho quan hệ r xác định trên tập thuộc tính $U=\{A_1,A_2,...,A_n\}$. Tập thuộc tính $X \subseteq U$. Phép chiếu quan hệ r lên tập thuộc tính X,

ký hiệu $\pi_x(r)$ được kết quả là một quan hệ xác định trên tập thuộc tính X gồm các bộ được lấy từ quan hệ r và có giá trị tại tập thuộc tính X.

Biểu diễn hình thức phép chiếu như sau:

$$\pi_x(r)=\{t[X]|\ t\in r\}$$

Nhận xét: phép chiếu thực chất là phép loại bỏ đi một số cột của quan hệ và giữ lại những cột còn lại của quan hệ đó.

1.1.4.5. Phép tích Đề các (Descartes)

Cho quan hệ r xác định trên tập thuộc tính $U=\{A_1,A_2,...,A_n\}$ và quan hệ s xác định trên tập thuộc tính $V=\{B_1,B_2,...,B_m\}$. Tích Đề các của hai quan hệ r và s, ký hiệu $r\times s$ là một quan hệ xác định trên tập thuộc tính $U\times V =\{ A_1,A_2,...,A_n, B_1,B_2,...,B_m)$ và được biểu diễn như sau:

$$r\times s = \{(t,u)|\ t\in r \text{ và } u\in s\}$$

1.1.4.6. Phép chọn (Selection)

Cho quan hệ r xác định trên tập thuộc tính $U=\{A_1,A_2,...,A_n\}$. Phép chọn từ quan hệ r các bộ t thỏa mãn biểu thức chọn F cũng là một quan hệ xác định trên U và được biểu diễn hình thức như sau:

$$\sigma_F(r)=\{t|\ t\in r \text{ và } F(t) \text{ đúng}\}$$

Trong đó, F(t) được hiểu là giá trị cuả các thuộc tính xuất hiện trong biểu thức s tại bộ t trả về giá trị đúng.

1.1.4.7. Phép kết nối (Join)

1.2. HỆ PHÂN TÁN

1.2.1. Đặc trưng của hệ phân tán

- Hệ tin học phân tán là hệ không chia sẻ bộ nhớ và đồng hồ.

- Phân tán hóa các quá trình xử lý và thực hiện các công việc đó trên các trạm xa nhau.
- Thời hạn truyền thông tin trong hệ không giống nhau, các thông điệp có thể bị mất trong quá trình chuyển tải, các thông điệp có thể được truyền kép và hệ thống có thể rơi vào sự cố.
- Một trạm nào đó bị sự cố thì không ảnh hưởng đến hệ thống, công việc của nó sẽ được phân cho các trạm khác.

1.2.2. Tính chất của hệ phân tán

1.2.2.1. Tính trong suốt

1.2.2.2. Tính hiệu quả

1.2.2.3. Tính mềm dẻo

1.2.2.4. Tính phù hợp

1.2.2.5. Tính bền vững

1.2.2.6. Tính mở

1.2.2.7. Tính song song

1.2.3. Các điểm mạnh trong hệ tin học phân tán

- Cơ chế tính toán phân tán hỗ trợ truy cập các dữ liệu được lưu ở nhiều nơi.
- Nhờ cơ chế nhân bản nên người dùng chỉ cần truy cập cục bộ cũng lấy được các thông tin từ các trung tâm chính ở rất xa.
- Nếu chúng ta không truy cập dữ liệu được tại vị trí này , chúng ta có thể thử ở nơi khác.
- Dữ liệu phân tán đòi hỏi phải được nhân bản và đồng bộ hóa cao thông qua các mối liên kết mạng , điều này làm cho việc quản trị và giám sát phức tạp hơn.

- Hệ phân tán được xây dựng trên giao thức TCP /IP và các kỹ thuật Web cùng với các ứng dụng trung gian (middleware) thúc đẩy việc tính toán phân tán.

1.2.4. Các mô hình ứng dụng Phân tán

1.2.4.1. Phân tán trên mạng cục bộ

1.2.4.2. Phân tán trên mạng diện rộng

CHƯƠNG 2

THIẾT KẾ VÀ TỐI ƯU HÓA TRUY VẤN PHÂN TÁN

2.1. THIẾT KẾ CƠ SỞ DỮ LIỆU PHÂN TÁN

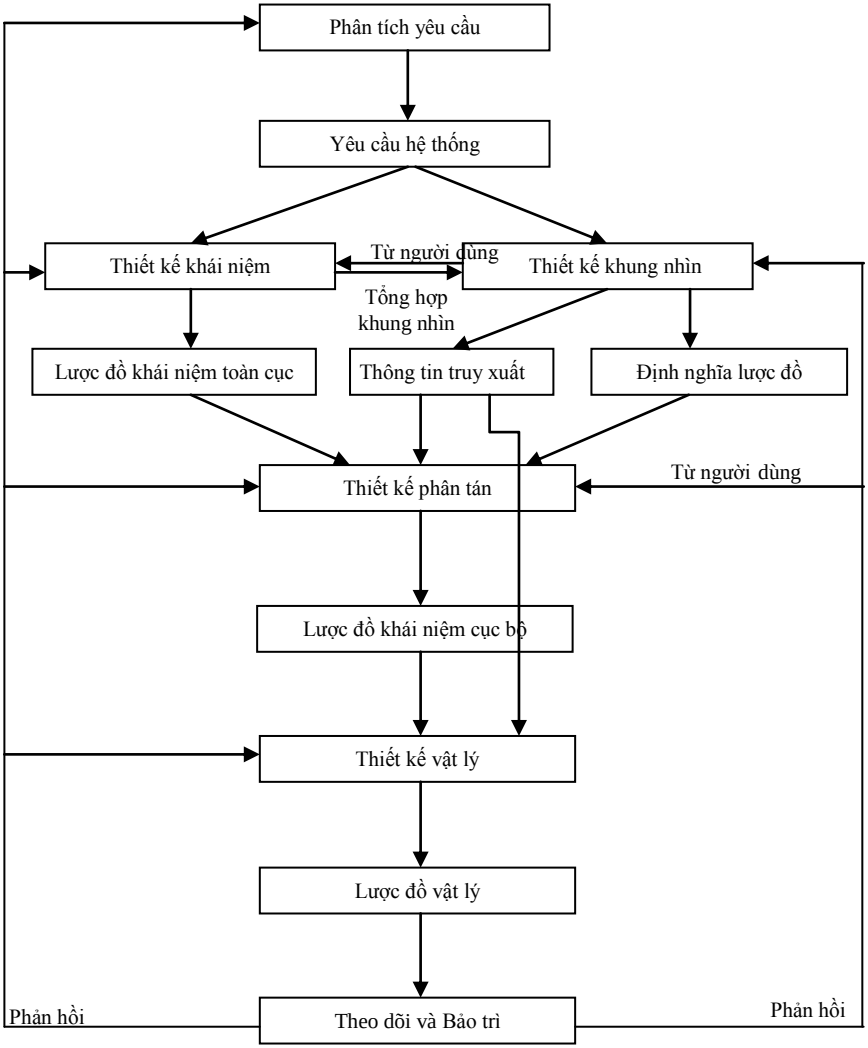
Thiết kế một hệ thống phân tán cần phải chọn những vị trí đặt dữ liệu và các chương trình trên một mạng máy tính. Đối với hệ quản trị CSDL phân tán, việc phân tán các ứng dụng đòi hỏi hai điều: phân tán hệ quản trị CSDL và phân tán các chương trình ứng dụng chạy trên hệ quản trị đó.

2.1.1. Các mục tiêu của thiết kế phân tán dữ liệu

- Tính cục bộ xử lý
- Tính sẵn sàng và độ tin cậy của dữ liệu phân tán
- Điều phối tải làm việc
- Các chi phí lưu trữ và khả năng lưu trữ có sẵn.

2.1.2. Các chiến lược thiết kế

2.1.2.1 Quá trình thiết kế từ trên xuống (top-down)



Hình 1.1. Quá trình thiết kế Top-Down

Thiết kế khung nhìn có nhiệm vụ định nghĩa các giao diện cho người sử dụng cuối.

Thiết kế khái niệm là quá trình xem xét tổng thể nhằm xác định các loại thực thể và mối liên hệ giữa các thực thể.

Lược đồ khái niệm toàn cục và thông tin về kiểm mẫu truy xuất thu được trong thiết kế khung nhìn sẽ là đầu vào cho bước thiết kế phân tán. Mục tiêu của giai đoạn này là thiết kế các lược đồ khái niệm cục bộ bằng cách phân tán các thực thể cho các vị trí của hệ thống phân tán.

Thiết kế phân tán gồm có hai bước: phân mảnh và cấp phát.

Bước cuối cùng là thiết kế vật lý, là bước ánh xạ lược đồ khái niệm cục bộ sang các thiết bị lưu trữ vật lý có sẵn tại các vị trí tương ứng. Đầu vào cho quá trình này là lược đồ khái niệm cục bộ va thông tin về kiểu mẫu truy xuất các mảnh.

2.1.2.2 Quá trình thiết kế từ dưới lên

Quá trình này thích hợp với những CSDL được thiết kế từ đầu. Tuy nhiên chúng ta rất hay gặp trong thực tế là đã có sẵn một số CSDL, và nhiệm vụ thiết kế là phải tích hợp chúng thành một CSDL.

2.1.3 Các vấn đề thiết kế phân tán

2.1.3.1 Các lý do phân mảnh

- Trước tiên, khung nhìn của các ứng dụng thường chỉ là một tập con của quan hệ;
- Thứ hai là nếu các ứng dụng có các khung nhìn được định nghĩa trên một quan hệ cho trước lại nằm tại những vị trí khác nhau thì có hai cách chọn lựa với đơn vị phân tán là toàn bộ quan hệ, hoặc quan hệ không được nhân bản mà được lưu ở một vị trí

hoặc quan hệ được nhân bản cho tất cả hoặc một số vị trí có chạy ứng dụng.

- Cuối cùng, việc phân rã một quan hệ thành nhiều mảnh, mỗi mảnh được xử lý như một đơn vị, sẽ cho phép thực hiện nhiều giao dịch đồng thời. Ngoài ra, việc phân mảnh các quan thệ sẽ cho phép thực hiện song song một câu vấn tin bằng cách chia nó thành một tập các câu vấn tin con hoạt tác trên các mảnh. Vì thế việc phân mảnh sẽ làm tăng mức độ hoạt động đồng thời và như thế làm tăng lưu lượng hoạt động của hệ thống. Kiểu hoạt động đồng thời này được gọi là đồng thời nội vấn tin.

2.1.3.2 Các kiểu phân mảnh

Có ba phương pháp khác nhau: chia bảng theo chiều dọc, chia bảng theo chiều ngang và chia bảng hỗn hợp.

2.1.3.3. Mức độ phân mảnh

Mức độ phân mảnh có thể là từ thái cực không phân mảnh nào đến thái cực phân mảnh thành từng bộ (trường hợp phân mảnh ngang) hoặc thành từng thuộc tính (trường hợp phân mảnh dọc).

2.1.3.4. Các quy tắc phân mảnh

- Tính đầy đủ
- Tính tái thiết được
- Tính tách biệt

2.1.3.5. Các kiểu cấp phát

Khi dữ liệu được cấp phát, nó có thể được nhân bản hoặc chỉ duy trì một bản duy nhất.

2.1.3.6 Các yêu cầu thông tin

Các thông tin cần cho thiết kế phân tán có thể phân chia thành bốn loại: thông tin CSDL, thông tin ứng dụng, thông tin về mạng và thông tin về hệ thống máy tính.

2.1.4 Phương pháp phân mảnh

Thiết kế phân mảnh bao gồm công việc nhóm các bộ trong trường hợp phân mảnh ngang hay nhóm các thuộc tính trong trường hợp phân đoạn dọc có cùng đặc tính theo quan điểm cấp phát.

2.1.4.1 Phân mảnh ngang

Phân mảnh ngang chia một quan hệ theo các bộ. Vì vậy mỗi mảnh là một tập con của quan hệ. Có hai loại phân mảnh ngang: phân mảnh ngang nguyên thuỷ và phân mảng ngang dẫn xuất.

2.1.4.2. Phân mảnh dọc

Một phân mảnh dọc cho một quan hệ r sinh ra các mảnh r_1, r_2, ..., r_n, mỗi mảnh chứa một tập con thuộc tính của R và cả khoá của r. Mục đích của phân mảnh dọc là phân hoạch một quan hệ thành một tập các quan hệ nhỏ hơn để nhiều ứng dụng có thể chỉ chạy trên một quan hệ.

2.1.4.2. Phân mảnh hỗn hợp

- Phân mảnh ngang cho các mảnh phân chia theo chiều dọc
- Phân mảnh dọc cho các mảnh phân chia theo chiều ngang.

2.1.5. Cấp phát cho các mảnh

Trong các công việc cấp phát cho các mảnh, quan trọng phân biệt được thiết kế cấp phát cho các mảnh dư thừa hay không dư thừa, cách dễ nhất là hướng "phù hợp nhất": tiêu chuẩn vị trí kết hợp với khả năng cấp phát cho các mảnh.

2.2. TỐI ƯU HÓA TRUY VẤN

2.2.1. Nguyên tắc tối ưu hoá

- Ưu tiên thực hiện các phép chiếu và chọn;
- Trước khi phải thực hiện phép tích Đề các, hãy tìm chiến lược truy nhập tốt nhất vào CSDL;
- Thực hiện các phép kết nối cân bằng chi phí sẽ rẻ hơn nhiều so với chi phí thực hiện phép tích Đề các;
- Nhóm các phép toán chọn và chiếu liên tiếp thành một phép toán duy nhất;
- Nhóm các phép tích và chiếu liên tiếp thành một phép toán duy nhất. Trong khi thi thực hiện phép tích có thể giới hạn chi phí thực hiện bằng phép chiếu;
- Tìm biểu thức chung trong một biểu thức;
- Đánh giá sơ bộ trước khi thực hiện truy vấn.

2.2.2. Tối ưu hoá các biểu thức đại số quan hệ

2.2.2.1. Biểu thức quan hệ

Biểu thức quan hệ là một biểu thức mà các toán hạng là các quan hệ trong một CSDL và các phép toán là các phép toán trong đại số quan hệ. Mỗi biểu thức quan hệ thông thường là một truy vấn của người sử dụng.

2.2.2.2. Biến đổi biểu thức quan hệ

Hai biểu thức đại số quan hệ được gọi là tương đương nếu trên một mô hình dữ liệu bất kỳ, hai biểu thức này tạo ra hai quan hệ có tập các bộ dữ liệu giống nhau.

2.2.2.3. Các quy tắc tương đương

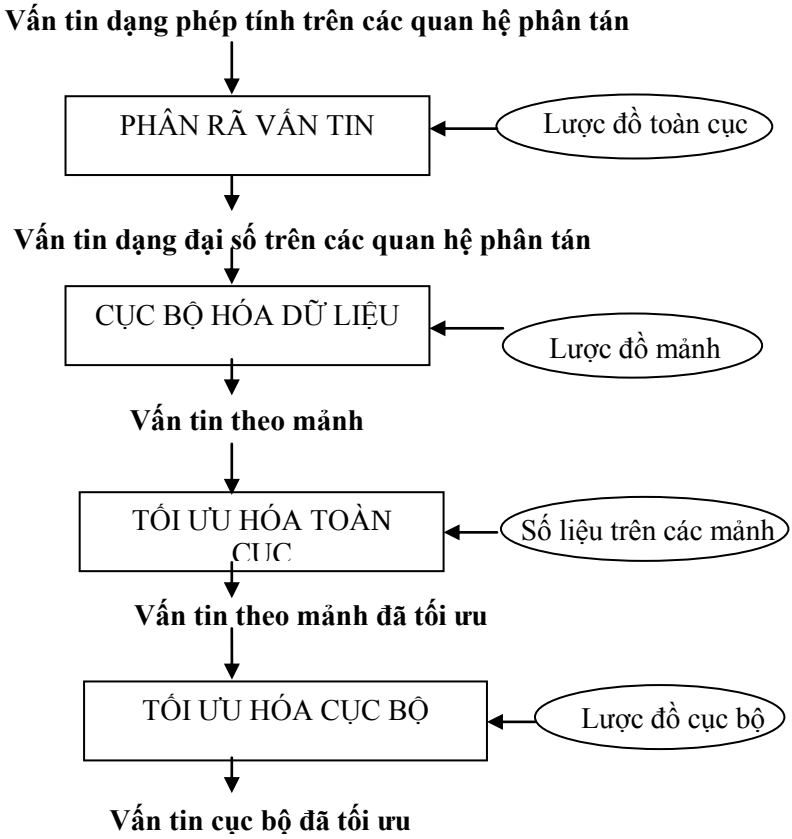
Hai biểu thức tương đương nếu thay thế một biểu thức của dạng thứ nhất bởi một biểu thức của dạng thứ hai, và ngược lại ta có thể thay thế biểu thức của dạng thứ hai bởi biểu thức của dạng thứ nhất thì hai biểu thức cùng tạo ra kết quả giống nhau trên bất kỳ hệ CSDL. Các quy tắc sau được sử dụng để biến đổi các biểu thức tương đương với nhau.

2.2.3. Mô hình tối ưu hóa trong môi trường phân tán

2.2.3.1. Mô hình tối ưu hóa trong môi trường tập trung

Câu truy vấn → Kiểm tra ngữ pháp → Truy vấn đúng ngữ pháp → Kiểm tra hợp lệ → Truy vấn SQL hợp lệ → Dịch truy vấn → Truy vấn đại số quan hệ → Tối ưu hóa đại số quan hệ → Truy vấn đại số quan hệ đã tối ưu → Chọn chiến lược đã tối ưu → Kế hoạch thực hiện → Tạo sinh mã → Mã truy vấn

2.2.3.2. Mô hình tối ưu hóa trong môi trường phân tán



Hình 2.1. Mô hình tối ưu hóa truy vấn phân tán

2.2.4. Chi phí

Total cost= CPU cost + I/O cost + communication cost

CPU cost= unit instruction cost * no.of instructions

I/O cost = unit disk I/O cost * no. of disk I/Os

communication cost = message initiation + transmission

Trong đó:

Total cost : Tổng chi phí

CPU cost : Chi phí CPU

communication cost : Chi phí truyền thông

unit instruction cost : Chi phí đơn vị lệnh

no.of instructions : Số câu lệnh

unit disk I/O cost : Chi phí đơn vị vào ra

no. of disk I/Os : Kích thước truy xuất vào ra trên đĩa

message initiation : Khởi tạo thông điệp

transmission : Truyền thông

2.2.5. Các thuật toán

2.2.5.1. Thuật toán D-Ingres

Hàm mục tiêu của thuật toán là làm giảm thiểu cho phí tổ hợp của thời gian truyền tin và thời gian đáp ứng, tuy nhiên hai mục tiêu này có thể xung khắc với nhau, thuật toán tối ưu hóa vấn tin bỏ qua chi phí truyền dữ liệu đến vị trí nhận kết quả. Thuật toán cũng tận dụng phân mảnh nhưng để cho đơn giản nó chỉ xét phân mảnh ngang.

Đầu vào cho thuật toán xử lý vấn tin là một câu vấn tin được diễn tả bằng phép tính quan hệ bộ (ở dạng chuẩn hội) và thông tin lược đồ (kiểu mạng, vị trí và kích thước của mỗi mảnh).

*2.2.5.2. Thuật toán System R**

Đầu vào cho thuật toán này là câu vấn tin đã cục bộ hóa được diễn tả như một cây vấn tin, vị trí của các quan hệ và số liệu thống kê của chúng.

Đầu ra là chiến lược có chi phí nhỏ nhất.

2.2.5.3. Thuật toán SDD-1

Đầu vào cho thuật toán gồm có đồ thị vấn tin *m* vị trí các quan hệ và số liệu thống kê về quan hệ. Sau khi hoàn tất xử lý cục bộ ban đầu, một giải pháp khả thi ban đầu được chọn ra, đó là một *lịch biểu thực thi toàn cục* (global execution schedule) có chứa tát cả các truyền thông giữa các vị trí. Kết quả này có được từ việc tính chi phí thấp nhất.

CHƯƠNG 3
XÂY DỰNG HỆ THỐNG QUẢN LÝ NHÂN VIÊN

3.1. PHÁT BIỂU BÀI TOÁN

Để áp dụng những lý thuyết đã nghiên cứu và áp dụng vào việc Tin học hoá quản lý hành chính Nhà nước trong Sở Khoa học và Công nghệ Bình Định, tôi chọn hướng phát triển quản lý nhân sự.

3.2. PHÂN TÍCH HỆ THỐNG CSDL

3.2.1. Các đối tượng tham gia vào hệ thống

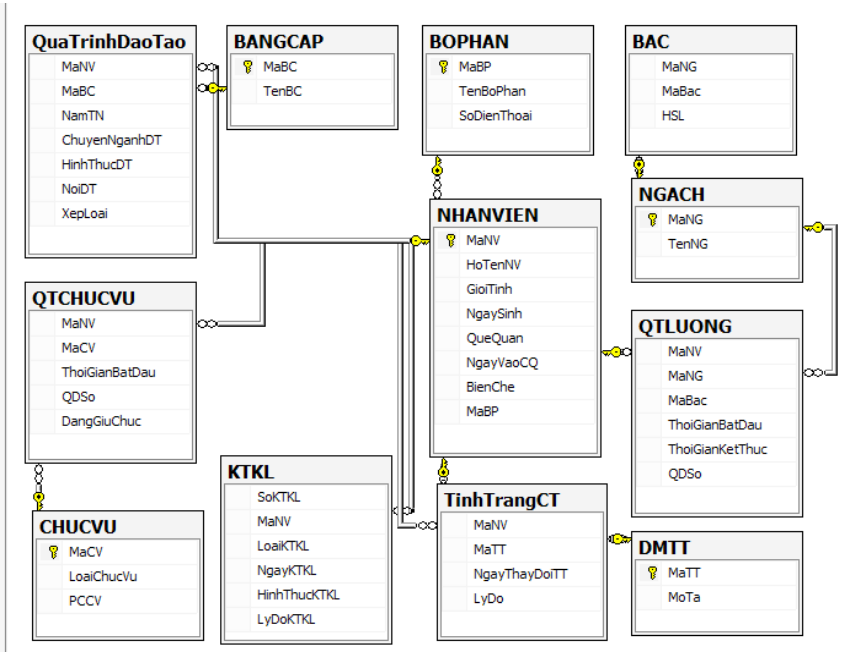
Ban giám đốc; Phòng Tổ chức; Phòng Tài vụ; Các Trung tâm trực thuộc; Cán bộ, nhân viên.

3.2.2. Yêu cầu về chức năng của hệ thống

- Cập nhật hồ sơ nhân viên;
- Cập nhật các biến động trong quá trình công tác của nhân viên;
- Phân tích và báo cáo tình hình nhân sự;
- Tìm kiếm, tra cứu thông tin về nhân sự theo một tiêu chí nào đó;
- Phân cấp hệ thống.

3.2.3. Xây dựng các thực thể và quan hệ của các thực thể

3.2.4. Mô hình cơ sở dữ liệu quan hệ



Hình 3.13. Mô hình quan hệ

3.2.5. Phát biểu, phân tích ứng dụng

3.2.5.1. Ví dụ 1

Xét câu truy vấn sau: " *Hãy cho biết Ngày vào cơ quan của nhân viên có mã là 24*"

Câu SQL như sau:

```
SELECT    NgayVaoCCQ
FROM      NHANVIEN
```

```
WHERE    (MaNV = '24')
```

Biểu thức đại số quan hệ: $\Pi_{NgayVaoCQ}(\sigma_{MaNV = '24'}(NHANVIEN))$

3.2.5.2. Ví dụ 2

Xét câu truy vấn sau: " *Hãy cho biết tên các nhân viên đang giữ chức vụ là Phó Phòng?*"

Câu SQL như sau:

```
SELECT NHANVIEN.HoTenNV
FROM  QTCHUCVU, NHANVIEN
WHERE (QTCHUCVU.MaCV = 'PP') AND
       (QTCHUCVU.DangGiuChuc = 'TRUE')AND
       (QTCHUCVU.MaNV = NHANVIEN.MaNV)
```

Biểu thức đại số quan hệ như sau:

$$\Pi_{HoTenNV}(\sigma_{(MaCV = 'PP') \wedge (DangGiuChuc='TRUE') \wedge (QTCHUCVU.MaNV = NHANVIEN.MaNV)}(QTCHUCVU*NHANVIEN))$$

Trước tiên sẽ tính tích Đề-các của hai quan hệ QTCHUCVU và NHANVIEN, sau đó mới chọn ra các bộ thỏa mãn điều kiện (MaCV = 'PP') $\wedge$ (DangGiuChuc='TRUE')$\wedge$ (QTCHUCVU.MaNV = NHANVIEN.MaNV) và cuối cùng sử dụng phép chiếu để lấy HoTenNV.

Áp dụng phép biến đổi tương đương (phép giao hoán giữa phép chọn và phép Đề các), chúng ta có biểu thức đại số quan hệ cho câu truy vấn mới:

$$\Pi_{HoTenNV}(\sigma_{(MaCV = 'PP') \wedge (DangGiuChuc='TRUE')}(QTCHUCVU) \triangleright\triangleleft NHANVIEN)$$

Với câu truy vấn này thì đã loại bỏ đi đáng kể các bộ không thỏa mãn yêu cầu trong tích Đề-các, mà chỉ chọn ra các bộ theo yêu

cầu và sau đó nối tự nhiên với quan hệ NHANVIEN. Câu truy vấn tối ưu hơn như sau:

```
SELECT   NHANVIEN.HoTenNV
FROM    QTCHUCVU INNER JOIN
      NHANVIEN ON QTCHUCVU.MaNV =
        NHANVIEN.MaNV
WHERE (QTCHUCVU.MaCV = 'PP')AND
       (QTCHUCVU.DangGiuChuc = 'TRUE')
```

3.2.5.3. Ví dụ 3

Cho biết Họ tên các nhân viên có phụ cấp chức vụ là 0.3?

Câu truy vấn SQL như sau:

```
SELECT NHANVIEN.HoTenNV
FROM NHANVIEN, QTCHUCVU, CHUCVU
WHERE(CHUCVU.PCCV= 0.3) AND
       (QTCHUCVU.DangGiuChuc = 'TRUE') AND
       (CHUCVU.MaCV = QTCHUCVU.MaCV) AND
       (QTCHUCVU.MaNV = NHANVIEN.MaNV)
```

Với câu truy vấn trên, mệnh đề WHERE ở dạng chuyển hội

```
(CHUCVU.PCCV= 0.3) AND (QTCHUCVU.DangGiuChuc =
'TRUE') AND (CHUCVU.MaCV = QTCHUCVU.MaCV) AND
(QTCHUCVU.MaNV = NHANVIEN.MaNV)
```

Hệ thống sẽ thực hiện tích Đề-Các quan hệ NHANVIEN, QTCHUCVU, CHUCVU rồi sau đó chọn ra các bộ thỏa mãn điều kiện.

Biểu thức đại số quan hệ như sau:

$$\Pi_{HoTenNV}(\sigma_{(PCCV = 0.3) \wedge (DangGiuChuc='TRUE') \wedge (CHUCVU.MaCV=QTCHUCVU.MaCV) \wedge (QTCHUCVU.MaNV =NHANVIEN.MaNV)}((CHUCVU* QTCHUCVU)*NHANVIEN))$$

Cũng thực hiện phép biến đổi tương đương, chúng ta có biểu thức đại số quan hệ mới như sau:

$$\Pi_{HoTenNV}(\Pi_{MaNV}(\Pi_{MaCV,MaNV}(\sigma_{PCCV = 0.3}(CHUCVU) \triangleright\triangleleft \sigma_{DangGiuChuc = 'TRUE'} QTCHUCVU) \triangleright\triangleleft NHANVIEN))$$

```
SELECT    NHANVIEN.HoTenNV
FROM      NHANVIEN INNER JOIN
            QTCHUCVU ON NHANVIEN.MaNV =
          QTCHUCVU.MaNV INNER JOIN
            CHUCVU ON QTCHUCVU.MaCV =
          CHUCVU.MaCV
WHERE    (CHUCVU.PCCV = 0.3) AND
           (QTCHUCVU.DangGiuChuc = 'TRUE')
```

3.2.6. Mô hình CSDL phân tán

Quá trình tạo ra các mảnh ở nhiều vị trí khác nhau, dựa vào:

- Khi tần suất truy vấn dữ liệu của nhiều nhóm người dùng thì chúng ta có thể xác định được đâu là nơi truy cập nhiều, và dữ liệu nào được sử dụng nhiều ở đó, như thế chúng ta mới có cơ sở để phân bố.
- Thao tác đối với CSDL là gì.

Ví dụ chúng ta xét ứng dụng truy vấn sau:

```
SELECT NHANVIEN.HoTenNV
FROM  NHANVIEN, QTCHUCVU
```

```
WHERE  NHANVIEN.MaNV = QTCHUCVU.MaNV)
        AND (QTCHUCVU.MaCV='PP')AND
(QTCHUCVU.DangGiuChuc='TRUE')
```

Biểu thức tương đương trong đại số quan hệ:

$$\Pi_{HoTenNV}(NHANVIEN \triangleright\triangleleft \sigma_{(MaCV='PP')AND(DangGiuChuc='TRUE')}(QTCHUCVU))$$

Giả sử quan hệ NHANVIEN và QTCHUCVU được phân mảnh:

$QTCHUCVU1 = \sigma_{MaNV <= 100}(QTCHUCVU)$

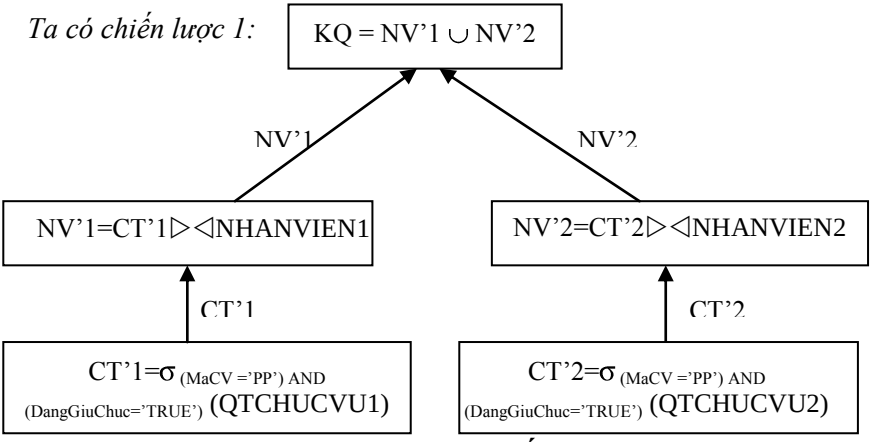
$QTCHUCVU2 = \sigma_{MaNV>100}(QTCHUCVU)$

$NHANVIEN1 = \sigma_{MaNV <= 100}(NHANVIEN)$

$NHANVIEN2 = \sigma_{MaNV >100}(NHANVIEN)$

Các mảnh QTCHUCVU1, QTCHUCVU2, NHANVIEN1, NHANVIEN2 được phân bố ở các vị trí trí tương ứng là 1, 2, 3 và 4. Mảnh Kết quả (KQ) được phân bố ở vị trí 5.

Ta có chiến lược 1:



Hình 3.14. Mô hình chiến lược 1

Chiến lược 2: quan hệ KQ được tính toán sau khi di chuyển các bộ của các mảnh về vị trí 5, cụ thể như sau:

$$KQ = (NHANVIEN1 \cup NHANVIEN2) \triangleright\triangleleft \sigma_{(MaCV='PP') \text{ AND } (DangGiuChuc='TRUE')} (QTCHUCVU1 \cup QTCHUCVU2)$$

Giả sử:

Thao tác truy xuất 1 bộ, ký hiệu là tupacc là 1 đơn vị

Thao tác truyền 1 bộ, ký hiệu là tuptrans là 10 đơn vị

Quan hệ NHANVIEN có 1000 bộ

Quan hệ QTCHUCVU có 400 bộ

Hiện nay có 20 trưởng khoa đang tại vị

Chi phí (đơn vị) của chiến lược 1 như sau:

(1) Tạo ra CT' cần:	(10+10)*tupacc =	20
(2) Truyền CT' NHANVIEN cần:	(10+10)*tuptrans=	200
(3) Tạo ra NV' cần:	(10+10)*tupacc*2 =	40
(4) Truyền NV' đến vị trí KQ cần:	(10+10)*tuptrans=	200
Tổng chi phí		460

Chi phí (đơn vị) của chiến lược 2 như sau:

(1) Truyền NHANVIEN đến KQ:	1000*tuptrans =	10000
(2) Truyền QTCHUCVU đến KQ:	400*tuptrans=	4000
(3) Tạo ra CT' cần:	400*tupacc =	400
(4) Nối NHANVIEN và CT' cần:	1000*20*tupacc=	20000
Tổng chi phí		34400

Đánh giá kết quả: việc di chuyển tất cả các bộ về KQ rồi tính toán sẽ tốn chi phí cao hơn. Từ đó, chúng ta chọn ra được chiến lược 1 để thực hiện.

Xác định yêu cầu thiết kế hệ thống quản lý nhân viên:

- Thông tin cập nhật nhanh nhất;
- Trả lời yêu cầu truy vấn trong khoảng thời gian ngắn nhất;
- Quá trình tạo ra báo cáo nhanh, đúng kỳ hạn. Thông tin của báo cáo chính xác.
- Giá cả để xây dựng hệ thống nhỏ nhất.

3.2.7. Lựa chọn vị trí đặt CSDL và phân nhóm người sử dụng

Đánh giá vị trí đặt cơ sở dữ liệu theo một số tiêu chuẩn sao cho vị trí đặt cơ sở dữ liệu tiện lợi nhất:

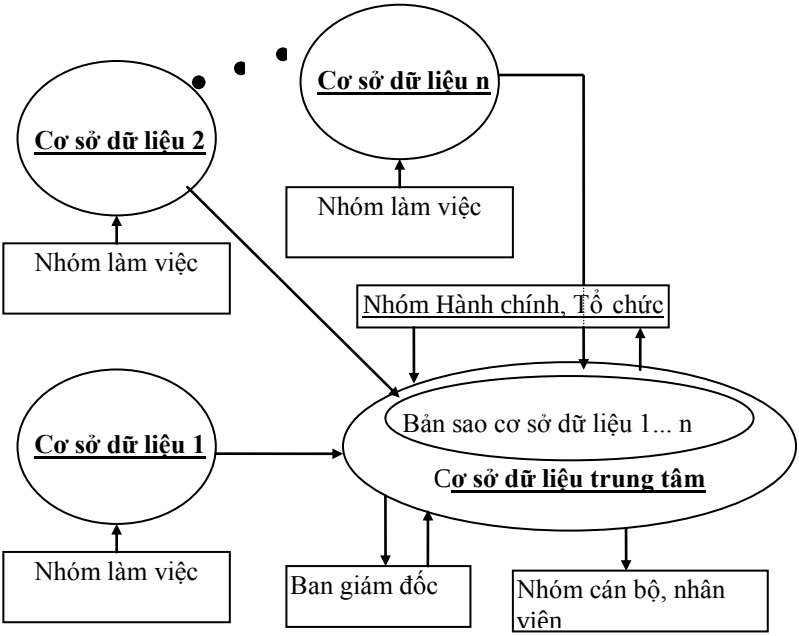
- Tần xuất sử dụng cơ sở dữ liệu;
- Số lần liên kết đến cơ sở dữ liệu;
- Các tham chiếu đến cơ sở dữ liệu để cập nhật, đọc hay thay đổi.

Dựa trên các tiêu chuẩn trên và tính chất của hệ thống thông tin quản lý nhân viên có thể lựa chọn hệ thống theo hai nhóm chính tương đương với hai cơ sở dữ liệu như sau:

Cơ sở dữ liệu 1: là nhóm người cập nhật thông tin vào hệ thống. Cơ sở dữ liệu của nhóm này đặt tại vị trí nhóm làm việc. Do yêu cầu của công việc, giữa các nhóm có thể lấy một số phần thông tin của nhau theo quyền.

Cơ sở dữ liệu trung tâm: gồm có nhóm là lãnh đạo (ban giám đốc), nhóm cán bộ nhân viên.

Sơ đồ mô tả cách kết nối thông tin giữa các cơ sở dữ liệu.



Hình 3.15. Mô hình cơ sở dữ liệu phân tán

Trong cơ sở dữ liệu, mỗi thực thể trong cơ sở dữ liệu là một bảng dữ liệu vật lý và các bảng vật lý này được hệ cơ sở dữ liệu quản lý và ngôn ngữ SQL trong chương trình sẽ quản lý việc cập nhật cơ sở dữ liệu này theo quyền.

Đánh giá kết quả thực hiện: Từ một cơ sở dữ liệu tập trung, thông qua các câu lệnh SQL, chúng ta sẽ dựa vào tần xuất truy cập cơ sở dữ liệu để thực hiện phân mảnh, các ứng dụng sẽ truy cập vào các mảnh để nâng cao hiệu quả sử dụng câu truy vấn. Với các mảnh ấy, chúng ta sẽ thực hiện cấp phát để tạo ra các bản sao cơ sở dữ liệu ở nhiều nơi, phục vụ cho nhiều đối tượng tham gia vào hệ thống, và kết quả tối ưu hóa đã giảm được chi phí khi truy vấn như thời gian thực hiện, chi phí truyền dữ liệu...

KẾT LUẬN

Những kết quả nghiên cứu của luận văn cho phép rút ra những kết luận sau:

Về mặt nghiên cứu lý thuyết: việc thiết kế cơ sở dữ liệu phân tán và quá trình tối ưu hóa câu truy vấn được phát triển từ hệ tập trung, việc phân tán các quan hệ thường chia chúng ra thành nhiều mảnh nhỏ hơn để đặt tại các vị trí thường xuyên sử dụng mảnh đó, các mảnh sau khi được chia ra sẽ cấp phát về những vị trí khác nhau. Vấn đề là làm sao để chúng ta có thể giảm thấp nhất được chi phí truy xuất, chi phí truyền thông, chi phí CPU đến mức thấp nhất mà vẫn đảm bảo được kết quả tương đương.

Luận văn cũng đã giới thiệu các thuật toán xử lý vấn tin như D-Ingress, System R* và SDD1.

Về mặt ứng dụng: áp dụng kỹ thuật tối ưu hóa phân tán vào thiết kế hệ thống quản lý nhân viên tại Sở Khoa học và Công nghệ Bình Định.

Hạn chế của luận văn là dừng lại ở mức độ mô tả, nghiên cứu lý thuyết thiết kế, tối ưu hóa và đưa ra mô hình ứng dụng chứ chưa đạt được ứng dụng cao nhất.

Trong thời gian tới, hướng phát triển của đề tài tập trung làm rõ hơn về phân tán dữ liệu và xây dựng ứng dụng phân tán cũng như cải tiến các thuật toán để đạt được kết quả cao hơn.