Cơ sở dữ liệu phân tán

Phạm Minh Khan
pmkhan@hcmunre.edu.vn

Chương 3: Thiết kế CSDL phân tán

Trong chương này có 2 vấn đề chính:

1. Một số tiêu chuẩn thiết kế về cách thức phân tán dữ liệu một cách hợp lý.
2. Nền tảng toán học hỗ trợ cho nhà thiết kế xác định sự phân tán dữ liệu

Chương 3: Thiết kế CSDL phân tán

Trong chương này giới thiệu cho chúng ta có 3 phần:

- Phần 1: Giới thiệu mô hình thiết kế cơ sở dữ liệu phân tán với hai tiếp cận từ trên xuống và từ dưới lên.
- Phần 2: Trình bày sự thiế kế phân mảnh ngang, phân mảnh dọc và phân mảnh hỗn hợp
- Phần 3: Trình bày sự cấp phát các phân mảnh, vấn đề này nhắm đến sự ánh xạ các phân mảnh đến các ảnh vậy lý.

Mô hình thiết kế cơ sở dữ liệu phân tán

➢Việc thiết kế cơ sở dữ liệu tập trung nhắm đến hai vấn đề sau:

1. Thiết kế lược đồ quan niệm
2. Thiết kế "cơ sở dữ liệu vật lý" nghĩa là ánh xạ lược đồ quan niệm đến các vùng lưu trữ và xác định các phương pháp truy xuất thích hợp.

Trong cơ sở dữ liệu phân tán hai vấn đề này trở thành vấn đề **thiết kế lược đồ phổ quát và thiết kế các cơ sở dữ liệu cục bộ tại mỗi site.**

Mô hình thiết kế cơ sở dữ liệu phân tán

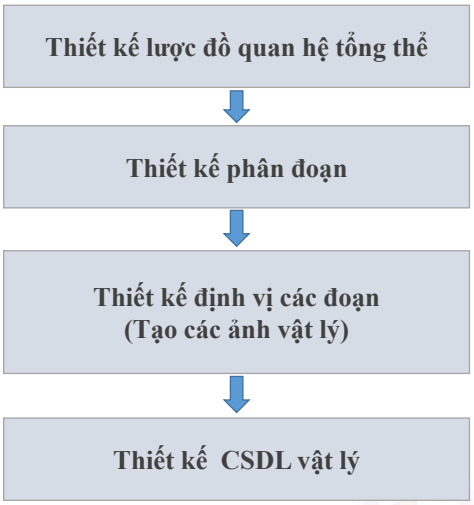
Sự phân tán cơ sở dữ liệu cộng thêm vào các vấn đề trên hai vấn đề mới:

- Thiết kế sự phân tán, nghĩa là xác định các quan hệ phổ quát được phân mảnh ngang, dọc hay hỗn hợp như thế nào?
- Thiết kế sự cấp phát các phân mảnh, nghĩa là xác định các phân mảnh được ánh xạ đến các ảnh vật lý như thế nào, kể cả việc xác định sự nhân bản dữ liệu.

➔Hai vấn đề mới này đặc trưng đầy đủ cho sự thiết kế phân tán dữ liệu

Mô hình thiết kế cơ sở dữ liệu phân tán

Sơ đồ thiết kế tổng thể



Mô hình thiết kế cơ sở dữ liệu phân tán

Việc thiết kế các chương trình ứng dụng được xây dựng sau khi thiết kế lược đồ, sự hiểu biết về các yêu cầu của các chương trình ứng dụng cũng quyết định đến sự thiết kế lược đồ vì các lược đồ phải hỗ trợ các ứng dụng một cách hiệu quả. Các yêu cầu của ứng dụng như sau:

- Site mà ứng dụng được đưa ra (còn được gọi là site gốc của ứng dụng)
- Tần số hoạt động của ứng dụng (nghĩa là số lượng yêu cầu hoạt động trong một đơn vị thời gian); trong trường hợp tổng quát các ứng dụng có thể được đưa ra từ nhiều sites, chúng ta cần biết tần số hoạt động của mỗi ứng dụng tại mỗi site.

Các mục tiêu của việc thiết kế phân tán dữ liệu

- **Sự truy xuất cục bộ** : Mục tiêu của sự phân tán dữ liệu là để các *ứng dụng truy xuất dữ liệu cục bộ càng nhiều càng tốt*, giảm bớt các truy xuất dữ liệu từ xa.
- **Tính sẵn sàng và khả tín (độ tin cậy) của các dữ liệu phân tán**: Nếu 1 quan hệ mà ta nhân bản thì tính sẵn sàng và khả tín là cao nhất, nếu 1 quan hệ mà nhân bản 1 phần (dữ liệu của nó nằm ít nhất ở 2 site) thì độ khả tín cũng cao, còn phân hoạch thì độ khả tín sẽ không tốt. ➔ Độ khả tín còn phụ thuộc vào quan hệ của ta đang ở góc độ là nhân bản hoàn toàn, nhân bản 1 phần hay phân hoạch.
- Sự phân bố tải: chắc là đạt, sự phân bố tải giúp chúng ta xử lý truy vấn dữ liệu sẽ nhanh hơn so với trên 1 cơ sở dữ liệu tập trung do ta có thể tận dụng được xử lý song song.

Các mục tiêu của việc thiết kế phân tán dữ liệu

- **Chi phí lưu trữ:** Sự phân tán cơ sở dữ liệu phản ánh chi phí của sự lưu trữ tại các sites khác nhau. Tuy nhiên chi phí lưu trữ dữ liệu không đáng kể so với chi phí xuất nhập, chi phí truyền thông của các ứng dụng. Nhưng giới hạn của bộ lưu trữ phải được xem xét kỹ.

➔**Ta có 4 tiêu chí, khi phân tán ta không thể đạt được hết 4 tiêu chí cùng 1 lúc được.**

Các hướng tiếp cận khi thiết kế sự phân tán dữ liệu

- **Hướng từ trên - xuống:**
  - Thiết kế lược đồ phổ quát
  - Thiết kế sự phân mảnh cơ sở dữ liệu
  - Cấp phát các mảnh đến các sites, tạo các ảnh vật lý của chúng

➔ Cách tiếp cận này nó rất phù hợp cho các hệ thống phát triển từ đầu (tức là nó chỉ có 1 db duy nhất) và nó cho phép thiết kế một cách hợp lý.

Các hướng tiếp cận khi thiết kế sự phân tán dữ liệu

- **Hướng từ dưới – lên:** Khi csdl phân tán được phát triển như là sự tổ hợp các cơ sở dữ liệu sẵn có thì nó không dễ dàng đối với PP tiếp cận từ trên - xuống. Từ đó cách tiếp cận từ dưới-lên có thể đợợc sử dụng để thiết kế sự phân tán dữ liệu. Cách thiết kế từ dưới lên yêu cầu:
  - Chọn một mô hình cơ sở dữ liệu chung để mô tả lược đồ phổ quát của cơ sở dữ liệu.
  - Chuyển dịch mỗi lược đồ cục bộ vào trong mô hình dữ liệu chung.
  - Tổ hợp lại lược đồ cục bộ vào trong lược đồ phổ quát chung.

Thiết kế sự phân mảnh dữ liệu

- Mục đích của việc thiết kế phân mảnh là **xác định các phân mảnh không chồng chéo lên nhau.**
- Thiết kế phân mảnh là nhóm các bộ (trong trường hợp phân mảnh ngang) hoặc nhóm các thuộc tính (theo phân mảnh dọc) mà có những tính chất giống nhau từ quan điểm cấp phát chúng. **Mỗi một nhóm các bộ hay các thuộc tính có cùng tính chất sẽ tạo nên một phân mảnh.**
- Khi ta tạo ra phân mảnh ngang thì những phân mảnh ngang đó thế nào là hợp lý thì phải thỏa 2 tính chất là **đầy đủ và cực tiểu**

Thiết kế sự phân mảnh dữ liệu

Xét 1 **Ví dụ 1**: ở csdl gốc ta có quan hệ tổng quát có tên là Employee thì khi ta phân tán phải đáp ứng được ứng dụng là yêu cầu thông tin từ quan hệ Employee về các nhân viên là thành viên của các dự án. **Mỗi phòng ban là một site của cơ sở dữ liệu phân tán.**

Tìm những dự án có sự tham gia của các nhân viên từ bất kỳ phòng ban nào, tất cả các site ?

Giải pháp:

Ưu tiên tìm các bộ nhân viên trong phòng ban đó trước với xác suất cao hơn ở những nhân viên của phòng ban khác.

Thiết kế sự phân mảnh dữ liệu

➢ Phân mảnh ngang nguyên thủy:

Cho R là quan hệ toàn cục mà chúng ta phân mảnh ngang nguyên thủy. Chúng ta đưa ra một số định nghĩa sau:

- Một **vị từ đơn giản** là vị từ có kiểu:

  Thuộc tính = giá trị
- Một **vị từ sơ cấp** Y cho một tập các vị từ đơn giản P là chuẩn hội của tất cả các vị từ xuất hiện trong P :

$$y = V (p^*_i)$$

Với $p^*_i = p_i$ hoặc p^*_i= not p_i và y = true

- Một phân mảnh là **một tập các bộ (dòng) tương ứng với một vị từ sơ cấp**
- Một **vị từ đơn giản p_i là thích hợp** đối với một tập các vị từ sơ cấp P **nếu tồn tại ít nhất hai vị từ sơ cấp** mà biểu thức của nó chỉ khác nhau do vị từ p_i (xuất hiện ở dạng thông thường và dạng phủ định của nó) mà các phân mảnh tương ứng được tham khảo đến bởi ít nhất một ứng dụng.

Thiết kế sự phân mảnh dữ liệu

Ví dụ 2: Xét sự phân mảnh ngang mà nó phức tạp hơn dựa vào 2 tiêu chí: 1 là DEPT (mã phòng ban) và JOB (công việc). Một ứng dụng yêu cầu các thông tin về các nhân viên tham gia vào các dự án (**Ví dụ 1**), **giờ lại có một ứng dụng quan trọng khác ngoài yêu cầu thông tin trên mà còn cần thông tin về nghề nghiệp.** Hai vị từ đơn giản cho ví dụ này là DEPT =1 và JOB =”P”. Các vị từ sơ cấp cho hai vị từ này là:

DEPT = 1 AND JOB = ”P”

DEPT = 1 AND JOB ≠ “P”

DEPT ≠ 1 AND JOB = “P”

DEPT ≠ 1 AND JOB ≠ “P”

Tất cả các vị từ đơn giản trên là thích hợp, trong khi, ví dụ, SAL > 50 không là một vị từ thích hợp;

Thiết kế sự phân mảnh dữ liệu

Cho P = {p_1, p_2, ..., p_n } là tập các vị từ đơn giản. Để P thể hiện sự **phân mảnh** một **cách đúng đắn và hiệu quả**. P phải **đầy đủ và cực tiểu**.

1. Tập các vị từ đơn giản P_r được gọi là **đầy đủ nếu và chỉ nếu xác suất mỗi ứng dụng truy xuất đến một bộ bất kỳ** thuộc về một mảnh hội sơ cấp nào đó được định nghĩa theo P_r **đều bằng nhau**.
2. **Tập các vị từ đơn giản** P_r được gọi là **cực tiểu nếu tất cả các vị từ của nó thích hợp** (nghĩa là **các phân mảnh** tương ứng được tham khảo đến bởi **ít nhất một ứng dụng**)

Thiết kế sự phân mảnh dữ liệu

Xét một ví dụ tổng quát:

Ví dụ gồm có các quan hệ **EMP, DEPT, SUPPLIER, SUPPLY**.

Giả sử cơ sở dữ liệu phân tán của công ty ở California có ba sites tại **San Francisco (site 1), Fresno (site 2), và Los Angeles (site 3);** Fresno nằm giữa San Francisco và Los Angeles.

Có tất cả **30 phòng** ban được nhóm lại như sau: **10 phòng ban đầu tiên ở gần San Francisco, các phòng ban từ 11 đến 20 ở gần Fresno và các phòng ban trên 20 thì ở gần Los Angeles.**

Tất cả các nhà cung cấp ở San Francisco hoặc ở Los Angeles. Ngoài ra công ty cũng được chia theo khái niệm miền: **San Francisco ở miền Bắc, Los Angeles ở miền nam** còn Fresno nằm giữa hai miền đó nên một số phòng ban nằm gần Fresno sẽ rơi vào miền bắc hoặc miền nam.

Thiết kế sự phân mảnh dữ liệu

Chúng ta thiết kế sự phân mảnh của SUPPLIER và DEPT với sự phân mảnh ngang nguyên thủy.

Các nhà cung cấp **trong quan hệ SUPPLIER**(SNUM, NAME, CITY) có giá trị của **thuộc tính CITY là "SF" hoặc là "LA".**

Giả sử có *một ứng dụng quan trọng yêu cầu cho biết tên nhà cung cấp (NAME) khi nhập mã số nhà cung cấp (SNUM)*. Câu lệnh SQL cho ứng dụng đó như sau:

```
Select NAME from SUPPLIER where SNUM = $X
```

Thiết kế sự phân mảnh dữ liệu

Ứng dụng được gọi tại bất kỳ site nào; nếu nó được gọi tại site 1, nó sẽ tham khảo đến SUPPLIERS có CITY = "SF" với xác suất 80%; nếu được gọi từ site 2, nó sẽ tham khảo đến SUPPLIERS của "SF" và "LA" với xác suất bằng nhau; nếu nó được gọi từ site 3, nó sẽ tham khảo đến SUPPLIERS của "LA" với xác suất 80%.

Điều này dẫn đến là các phòng ban sẽ liên hệ đến các nhà cung cấp ở gần đó.

Chúng ta đưa các vị từ sau:

p_1 : CITY = "SF"

P_2: CITY = "LA"

Tập $\{p_1, p_2\}$ là đầy đủ và cực tiểu.

Thiết kế sự phân mảnh dữ liệu

Mặc dầu đơn giản, ví dụ này minh họa hai tính chất quan trọng sau:

- Các vị từ thích hợp mô tả cho phân mảnh này không thể được suy ra bằng cách phân tích mã lệnh của ứng dụng.
- Quan hệ mật thiết giữa các vị từ giảm đi số lượng phân mảnh. Trong trường hợp này chúng ta nên xem xét những vị từ tương ứng với các vị từ sơ cấp sau:

  y_1: (CITY = "SF") AND (CITY = "LA")

  y_2: (CITY = "SF") AND NOT(CITY = "LA")

  y_3: NOT(CITY = "SF") AND (CITY = "LA")

  y_4: NOT(CITY = "SF") AND NOT(CITY = "LA")

Nhưng chúng ta đã biết rằng:

(CITY = "LA") NOT (CITY = "SF") và (CITY = "SF") NOT (CITY = "LA")

Vì thế chúng ta suy ra y_1 và y_4 mâu thuẫn lẫn nhau và y2 và y3 sẽ đơn giản thành hai vị từ p1 và p2.

Thiết kế sự phân mảnh dữ liệu

Bây giờ chúng ta hãy xét quan hệ phổ quát sau:

DEPT(DEPTNUM, NAME, AREA, MGRNUM)

Chúng ta sẽ tập trung vào các ứng dụng quan trọng sau:

Phân tán DB này thỏa mãn ứng dụng quản trị các phòng ban ở miền bắc được gọi tại site 1, ứng dụng quản trị các phòng ban ở miền nam thì gọi ở site 3.

Chúng ta đưa ra các vị từ sau:

p_1: DEPTNUM <= 10
p_2: 10 < DEPTNUM <= 20
p_3: DEPTNUM > 20
p_4: AREA = “North”
p_5: AREA = “South”

Thiết kế sự phân mảnh dữ liệu

Xây dựng các vị từ sơ cấp

Có một số quan hệ giữa các vị từ trên nhoị AREA= “North” kéo theo DEPTNUM > 20 là sai; vì thế sự phân mảnh giảm còn 4 phân mảnh:

y_1: DEPTNUM <=10
y_2: (10 < DEPTNUM <= 20) AND (AREA = “North”)
y_3: (10 < DEPTNUM <= 20) AND (AREA = “South”)
y_4: DEPTNUM > 20

Thiết kế sự phân mảnh dữ liệu

Xây dựng các vị từ sơ cấp

	p_4: AREA = "North"	p_5: AREA = "South"
p_1: DEPTNUM <= 10 p_2: 10 < DEPTNUM <= 20 p_3: DEPTNUM > 20	y_1	FALSE
	y_2	y_3
	FALSE	y_4

Thiết kế sự phân mảnh dữ liệu

- Sự phân mảnh dẫn xuất ngang: Phân mảnh ngang dẫn xuất dựa vào cái quan hệ khác
- Sự phân mảnh dọc
- Sự phân mảnh hỗn hợp

Sự cấp phát các phân mảnh

Khi ta tạo lược đồ phân mảnh gồm nhiều publication để đưa qua các site

Bước tiếp cận thứ 1: nếu publication đưa qua 1 site duy nhất thì mối ánh xạ 1 – 1 thì có ưu điểm cơ chế đồng bộ sẽ nhanh, tiết kiệm chi phí lưu trữ nhưng **có hạn chế không an toàn**.

Bước tiếp cận thứ 2: ánh xạ 1 – n: 1 publication qua 2 site tức là có 2 ảnh vật lý (2 subcription) cùng chung 1 publication thì tốn chi phí lưu trữ và thời gian đồng bộ ➔ an toàn dữ liệu

Tài liệu tham khảo

- Tài liệu giảng dạy cơ sở dữ liệu phân tán của PIIT
- Cơ sở dữ liệu phân tán, PGS.TS Nguyễn Mậu Hân

Thank you for listening