Chương 1 Lý thuyết về cơ sở dữ liệu phân tán

Mục tiêu của phần này là trình bày khái quát về lý thuyết cơ sở dữ liệu phân tán.

Phần này sẽ đề cập đến các vấn đề sau:

- *Giới thiệu về cơ sở dữ liệu phân tán:*

  Giới thiệu các khái niệm về cơ sở dữ liệu phân tán, so sánh cơ sở dữ liệu phân tán với cơ sở dữ liệu tập trung từ đó rút ra những lý do để phát triển một hệ thống dựa trên cơ sở dữ liệu phân tán, cuối cùng trình bày khái quát về hệ quản trị cơ sở dữ liệu phân tán.

- *Kiến trúc về cơ sở dữ liệu phân tán*

- *Kiến trúc của hệ quản trị cơ sở dữ liệu phân tán.*

1.1 Giới thiệu về cơ sở dữ liệu phân tán

Trong những năm gần đây, cơ sở dữ liệu phân tán đã trở thành một lĩnh vực xử lý thông tin quan trọng và chúng ta dễ dàng nhận ra tầm quan trọng của nó ngày càng lớn mạnh. Chúng ta có lý do về tổ chức cũng như về kỹ thuật để phát triển theo xu hướng này: cơ sở dữ liệu phân tán loại đi nhiều thiếu sót của cơ sở dữ liệu tập trung và phù hợp hơn qua các cấu trúc phi tập trung cùng với các ứng dụng phân tán.

Chúng ta có thể định nghĩa sơ nét về cơ sở dữ liệu phân tán như sau: Một cơ sở dữ liệu phân tán là một tập hợp dữ liệu của một hệ thống nhưng được phân bố trên nhiều địa điểm (site) của một mạng máy tính. Định nghĩa này nhấn mạnh hai khía cạnh quan trọng của cơ sở dữ liệu phân tán là :

1. **Sự phân tán**: dữ liệu không lưu trữ trên cùng một địa điểm vì thế chúng ta có thể phân biệt nó với cơ sở dữ liệu tập trung.
2. **Mối tương quan luận lý** : Các dữ liệu có một số thuộc tính ràng buộc với nhau từ các cơ sở dữ liệu cục bộ mà được lưu trữ tại các địa điểm khác nhau trên mạng.

Ví dụ :

Xét một ngân hàng có ba chi nhánh nằm ở ba nơi khác nhau (hình 1.1). Tại mỗi chi nhánh, một hệ thống máy tính điều khiển các trạm thu hay rút tiền và quản lý cơ sở dữ

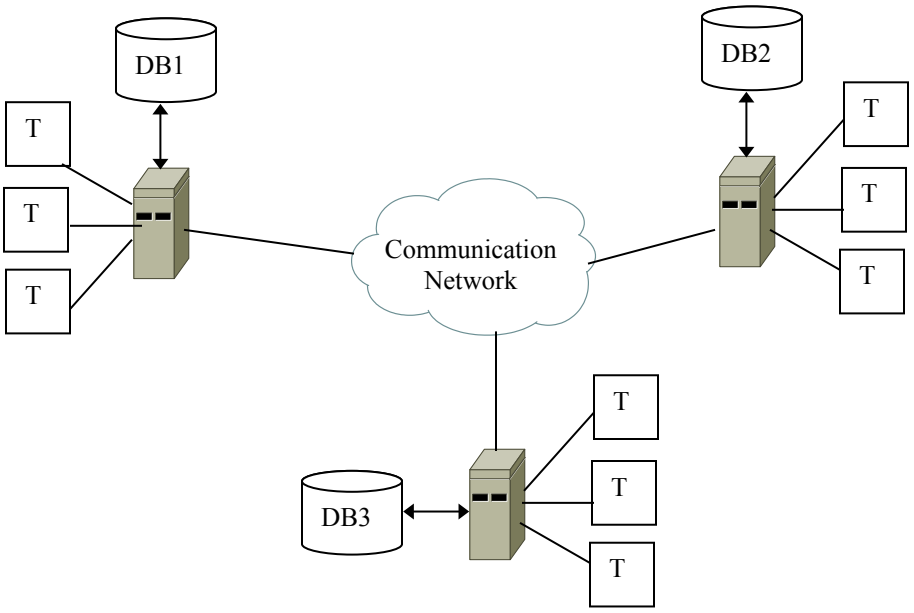
liệu về tài khoản. Mỗi hệ thống này với cơ sở dữ liệu tài khoản cục bộ tạo thành một site của cơ sở dữ liệu phân tán. Các hệ thống máy tính này được kết nối bởi một mạng truyền thông. Với những hoạt động thông thường, các yêu cầu từ các trạm chỉ cần truy xuất đến cơ sở dữ liệu tại chi nhánh của chúng. Vì thế ứng dụng này được gọi là ứng dụng cục bộ.

Với ví dụ trên nảy sinh hai câu hỏi sau: mỗi nhánh chỉ lưu trữ cơ sở dữ liệu cục bộ có đủ chưa? Cơ sở dữ liệu phân tán có phải là một tập các cơ sở dữ liệu cục bộ?

Để trả lời các câu hỏi này chúng ta tìm hiểu xem việc xử lý trên cơ sở dữ liệu cục bộ khác gì trên cơ sở dữ liệu phân tán. Về mặt kỹ thuật, chúng ta thấy cần có các ứng dụng mà truy xuất dữ liệu đang đặt ở nhiều nhánh. Các ứng dụng này được gọi là ứng dụng toàn cục hay ứng dụng phân tán.

Một ứng dụng toàn cục thông thường trong ví dụ trên là việc chuyển tiền từ một tài khoản này đến tài khoản khác. Ứng dụng này yêu cầu cập nhật cơ sở dữ liệu ở cả hai nhánh.

Hơn nữa ứng dụng toàn cục giúp cho người sử dụng không phân biệt được dữ liệu đó cục bộ hay từ xa. Đó là tính trong suốt dữ liệu trong cơ sở dữ liệu phân tán. Và đương nhiên khi ứng dụng toàn cục truy cập dữ liệu cục bộ sẽ nhanh hơn ứng dụng từ xa điều này nói lên sự nhân bản dữ liệu ở các nơi cũng làm tăng tốc độ xử lý chương trình.



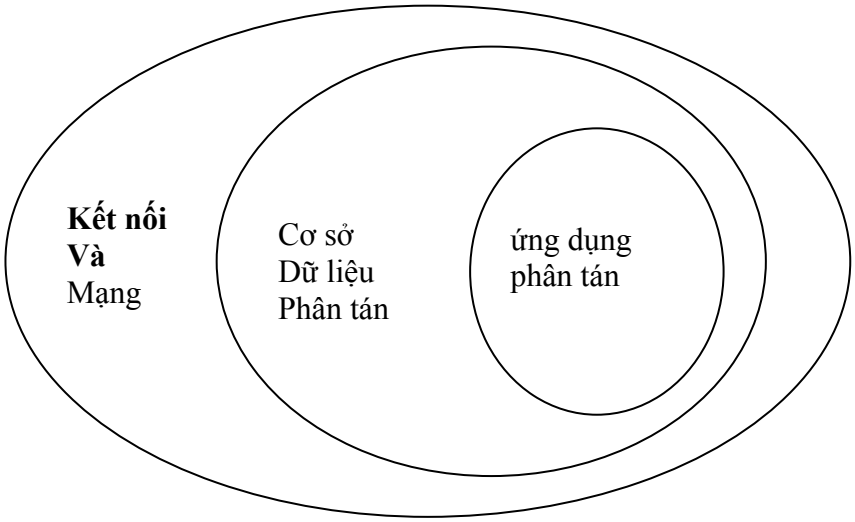
Hình 1.1 Cơ sở dữ liệu phân tán của ngân hàng có ba chi nhánh

1.1.1 **Định nghĩa về cơ sở dữ liệu phân tán**

Một cơ sở dữ liệu phân tán là tập hợp dữ liệu quan hệ lẫn nhau một cách luận lý trên cùng một hệ thống *nhưng được trãi rộng trên nhiều vị trí của một mạng máy tính.*

Mỗi vị trí có quyền tự quản cơ sở liệu cục bộ của mình và thực thi các ứng dụng cục bộ. Mỗi vị trí cũng phải ***tham gia vào việc thực thi ít nhất một ứng dụng toàn cục****: yêu cầu truy xuất dữ liệu tại nhiều vị trí qua mạng.*

Hình ảnh của cơ sở dữ liệu phân tán (hình 1.2) minh họa mối quan hệ của cơ sở dữ liệu phân tán với môi trường kết nối mạng máy tính và các ứng dụng phân tán.



Hình 1.2 Mối liên hệ giữa mạng máy tính, cơ sở dũ liệu phân tán và ứng dụng phân tán

1.1.2 Các điểm đặc trưng của cơ sở dữ liệu phân tán so với cơ sở dữ liệu tập trung

Cơ sở dữ liệu phân tán không đơn giản là việc phân tán các cơ sở dữ liệu tập trung bởi vì nó cho phép thiết kế các hệ thống có các tính chất khác với hệ thống tập trung truyền thống. Vì thế nên xem lại các tính chất đặc trưng của cơ sở dữ liệu tập trung truyền thống và so sánh nó với các tính chất của cơ sở dữ liệu phân tán. Các tính chất đặc trưng của cơ sở dữ liệu tập trung là điều khiển tập trung, độc lập dữ liệu, chuẩn hóa để loại bỏ sự dư thừa dữ liệu, các cấu trúc lưu trữ vật lý phức tạp đáp ứng cho việc truy xuất hiệu quả, toàn vẹn, phục hồi, điều khiển đồng thời và an toàn.

Dưới đây là bảng so sánh các tính chất đặc trưng của cơ sở dữ liệu tập trung và cơ sở dữ liệu phân tán:

Tính chất đặc trưng	**Cơ sở dữ liệu tập trung**	**Cơ sở dữ liệu phân tán**
Điều khiển tập trung	- Khả năng cung cấp sự điều khiển tập trung trên các tài nguyên thông tin. - Cần có người quản trị cơ sở dữ liệu.	- Cấu trúc điều khiển phân cấp: quản trị cơ sở dữ liệu toàn cục và quản trị cơ sở dữ liệu cục bộ phân tán.
Độc lập dữ liệu	- Tổ chức dữ liệu trong suốt với các lập trình viên. Các chương trình được viết có cái nhìn "quan niệm" về dữ liệu. - Lợi điểm: các chương trình không bị ảnh hưởng bởi sự thay đổi tổ chức vật lý của dữ liệu	- Ngoài tính chất độc lập dữ liệu như trong cơ sở dữ liệu tập trung, còn có tính chất trong suốt phân tán nghĩa là các chương trình được viết như cơ sở dữ liệu không hề được phân tán.
Sự dư thừa dữ liệu	Giảm thiểu sự dư thừa dữ liệu do: - Tính nhất quán dữ liệu cao - Tiết kiệm dung lượng nhớ.	- Giảm thiểu sự dư thừa dữ liệu đảm bảo tính nhất quán. - Nhưng lại nhân bản dữ liệu đến các địa điểm mà các ứng dụng cần đến, giúp cho việc thực thi các ứng dụng không dừng nếu có một địa điểm bị hỏng. Từ đó vấn đề quản lý nhất quán dữ liệu sẽ phức tạp hơn.
Các cấu trúc vật lý phức tạp và truy xuất hiệu quả	Các cấu trúc vật lý phức tạp giúp cho việc truy xuất dữ liệu được hiệu quả.	Các cấu trúc vật lý phức tạp giúp liên lạc dữ liệu trong cơ sở dữ liệu phân tán .
Tính toàn vẹn, phục hồi, đồng thời	Dựa vào giao tác.	Dựa vào giao tác phân tán.

Từ bảng so sánh trên, chúng ta thấy việc chọn lựa cơ sở dữ liệu phân tán sẽ thích hợp hơn đối với các ứng dụng phát triển trong một hệ thống mạng diện rộng do giảm được chi phí truyền thông để truy xuất dữ liệu.

1.1.3 Tại sao cần có cơ sở dữ liệu phân tán ?

1.1.3.1 Lý do tổ chức kinh tế

Nhiều tổ chức có cơ cấu tổ chức phi tập trung nên giải pháp cơ sở dữ liệu phân tán thích hợp hơn. Những năm gần đây do sự phát triển mạnh mẽ của công nghệ máy tính cùng với sự phát triển rộng rãi của các tổ chức kinh tế trên thế giới nên việc lưu trữ thông tin trên cơ sở dữ liệu tập trung cần xem xét lại về mặt hiệu quả.

1.1.3.2 Lý do kết nối các cơ sở dữ liệu hiện có

Cơ sở dữ liệu phân tán là giải pháp tự nhiên khi tổ chức đã có sẵn các cơ sở dữ liệu và cần mở rộng nó cho các ứng dụng phổ quát hơn. Trong trường hợp này cơ sở dữ liệu phân tán được xây dựng theo phương pháp từ dưới lên, dựa trên các cơ sở dữ liệu cục bộ có sẵn. Quá trình này có thể yêu cầu cấu trúc lại cơ sở dữ liệu cục bộ tuy nhiên công việc này lại đơn giản hơn xây dựng một cơ sở dữ liệu tập trung hoàn toàn mới.

1.1.3.3 Lý do tăng trưởng tổ chức

Nếu một tổ chức phát triển bằng cách thêm vào những đơn vị tổ chức tự quản như chi nhánh, kho bãi thì cách tiếp cận theo cơ sở dữ liệu phân tán hỗ trợ cho việc tăng trưởng cơ sở dữ liệu với mức độ ảnh hưởng nhỏ nhất. Trong khi đó cách tiếp cận theo cơ sở dữ liệu tập trung thì ngay từ đầu phải quan tâm đến sự phát triển của nó trong tương lai mà việc này thì khó dự đoán và tốn kém, nếu không dự liệu trước thì sẽ gây ra hậu quả nghiêm trọng không chỉ cho những ứng dụng mới mà còn cho cả hệ thống có sẵn.

1.1.3.4 Lý do tải truyền thông

Với một hệ cơ sở dữ liệu phân tán về mặt địa lý thì các ứng dụng truy cập sẽ giảm chi phí truyền thông so với cơ sở dữ liệu tập trung.

1.1.3.5 Đánh giá về hiệu suất

Sự tồn tại của các bộ xử lý tự quản nâng hiệu suất lên nhờ mức độ xử lý song song. Cơ sở dữ liệu phân tán có ưu thế là phân tán dữ liệu tại các địa điểm nên các ứng dụng có thể chạy riêng rẽ trên từng địa điểm và sự giao tiếp giữa các bộ xử lý là nhỏ nhất.

1.1.3.6 Độ tin cậy và tính hiệu quả

Mặc dầu việc phân tán dữ liệu làm tăng việc dư thừa dữ liệu trên toàn hệ thống nhưng lại cho chúng ta độ tin cậy và tính hiệu quả cao hơn trong cơ sở dữ liệu tập trung. Tuy nhiên để đạt được mục tiêu trên không phải dễ dàng mà đòi hỏi các kỹ thuật khá phức tạp. Sự hỏng hóc trong cơ sở dữ liệu phân tán có thể xảy ra thường hơn trong cơ sở dữ liệu tập trung vì số địa điểm tăng lên nhưng không bao giờ ảnh hưởng lên toàn hệ thống bởi thế nên nó có độ tin cậy và tính hiệu quả cao hơn cơ sở dữ liệu tập trung.

1.1.3.7 So sánh ưu và nhược điểm của việc phân tán dữ liệu

- Ưu điểm
  - Chia sẽ dữ liệu và điều khiển phân tán: Người sử dụng tại một vị trí này có thể truy xuất dữ liệu (được phép) ở vị trí khác. Hơn nữa việc quản trị cơ sở dữ liệu có thể được phân tán và thực hiện tự quản tại mỗi vị trí.
  - Độ tin cậy và tính sẵn sàng: Nếu một vị trí bị hỏng thì các vị trí còn lại trong hệ thống cơ sở dữ liệu phân tán vẫn tiếp tục hoạt động. Nếu dữ liệu được nhân bản ở một số vị trí thì một giao dịch cần truy xuất một mục dữ liệu có

thể tìm thấy ở bất kỳ vị trí nào trong số vị trí đó. Như thế sự cố tại một vị trí không ảnh hưởng đến hệ thống.

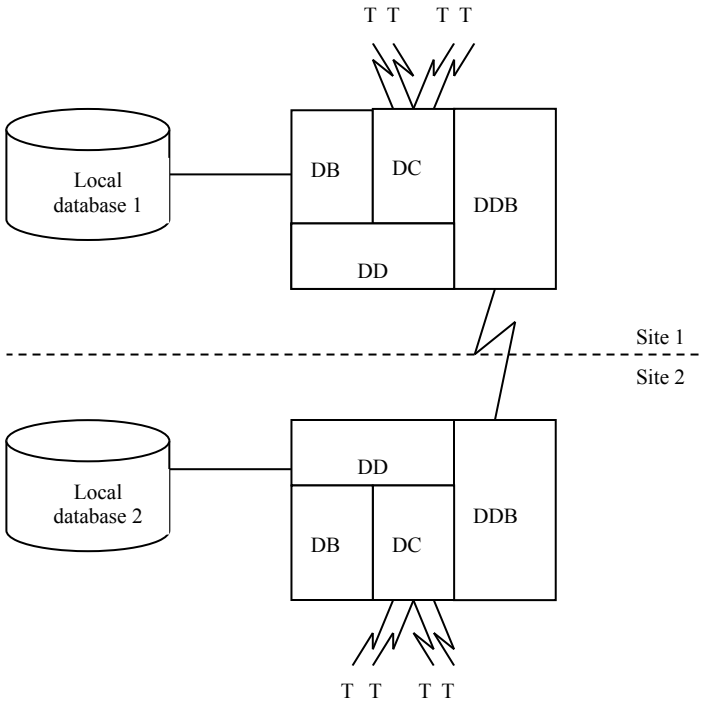
- Tăng tốc độ xử lý truy vấn: Nếu một truy vấn cần dữ liệu ở một số vị trí thì có thể chia câu truy vấn đó thành các câu truy vấn con rồi thực thi nó song song tại các vị trí.

- Nhược điểm
  - Chi phí phát triển phần mềm: Việc phát triển một hệ thống cơ sở dữ liệu phân tán khá phức tạp vì thế cần chi phí lớn.
  - Khó phát hiện lỗi: Việc phát hiện lỗi và đảm bảo tính đúng đắn của các thuật toán song song sẽ rất khó khăn.
  - Chi phí xử lý tăng: Sự trao đổi các thông báo và xử lý phối hợp giữa các vị trí sẽ tăng chi phí xử lý hơn trong các hệ thống tập trung.

1.1.4 Hệ quản trị cơ sở dữ liệu phân tán

Hệ quản trị cơ sở dữ liệu phân tán hỗ trợ việc tạo và duy trì cơ sở dữ liệu phân tán. Các hệ quản trị cơ sở dữ liệu phân tán hiện nay được phát triển bởi các nhà sản xuất các hệ quản trị cơ sở dữ liệu tập trung. Chúng chứa các thành phần bổ sung mở rộng các khả năng của các hệ quản trị cơ sở dữ liệu tập trung như hỗ trợ sự truyền thông và sự cộng tác giữa các hệ quản trị cơ sở dữ liệu trên các địa điểm khác nhau qua mạng máy tính. Các thành phần cơ bản cần thiết cho việc xây dựng một cơ sở dữ liệu phân tán là :

1. Thành phần quản trị cơ sở dữ liệu (DB Database Management)
2. Thành phần truyền dữ liệu (DC Data Communication)
3. Tự điển dữ liệu (DD Data Dictionnary) mở rộng để biểu diễn thông tin về sự phân tán dữ liệu trên mạng.
4. Thành phần cơ sở dữ liệu phân tán (DDB Distributed Database)



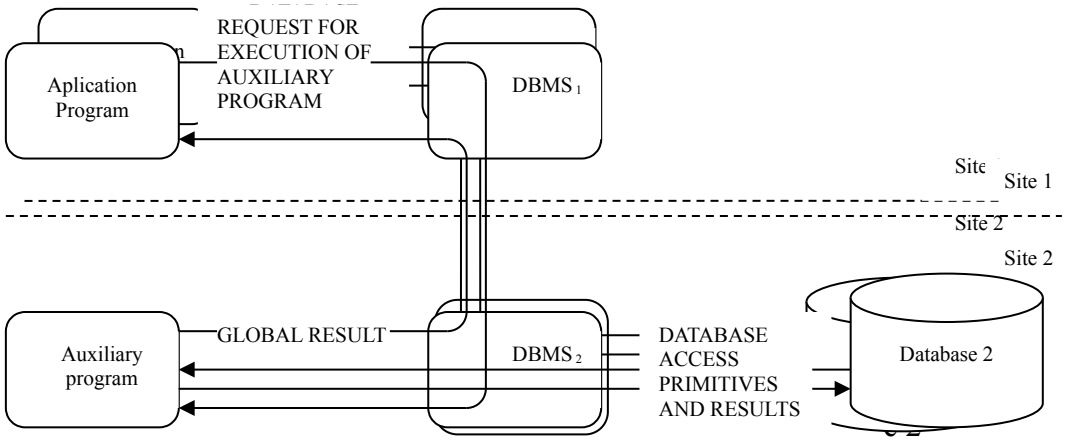
Hình 1.3 Các thành phần của hệ quản trị cơ sở dữ liệu

Các thành phần này được minh họa ở hình 1.3 đối với hai địa điểm trên mạng.

Các dịch vụ được hỗ trợ cho hệ thống trên thông thường là:

- Dịch vụ truy xuất cơ sở dữ liệu từ xa: tính chất này là một tính chất quan trọng nhất và được cung cấp bởi tất cả các hệ thống có thành phần cơ sở dữ liệu phân tán.
- Mức độ trong suốt của sự phân tán: tính chất này được hỗ trợ bởi các hệ thống khác nhau vì đó là sự cân bằng các yếu tố để đạt được sự kết hợp tốt nhất giữa sự trong suốt phân tán và hiệu suất.
- Hỗ trợ việc quản trị và điều khiển cơ sở dữ liệu: tính chất này bao gồm các công cụ để giám sát cơ sở dữ liệu, lấy thông tin về việc sử dụng cơ sở dữ liệu, cung cấp một cái nhìn toàn cục về các file dữ liệu lưu trữ trên các vị trí khác nhau.
- Hỗ trợ cho việc điều khiển đồng thời và phục hồi các giao tác phân tán.

(a) Truy xuất từ xa qua các lệnh có sẵn của hệ quản trị cơ sở dữ liệu



(b) Truy xuất từ xa qua chương trình bổ trợ

Hình 1.4 Các kiểu truy xuất đến cơ sở dữ liệu phân tán

Việc truy xuất đến một cơ sở dữ liệu từ xa bởi một ứng dụng có thể được thực hiện bởi một trong hai cách cơ bản minh họa ở hình 1.4. Hình 1.4a minh họa một ứng dụng đưa ra một yêu cầu tham khảo dữ liệu từ xa. Yêu cầu này được định tuyến bởi hệ quản trị cơ sở dữ liệu phân tán đến vị trí mà dữ liệu đó được lưu trữ, sau đó yêu cầu được thực thi tại vị trí đó và trả kết quả về. Trong cách này, đơn vị cơ bản liên lạc giữa các hệ thống là các nghi thức truy xuất cơ sở dữ liệu và kết quả nhận về cũng từ nghi thức này. Nếu các tiếp cận này được sử dụng cho việc truy xuất từ xa, sự trong suốt phân tán có thể được thực hiện bằng cách cung cấp các tên file toàn cục; các nghi thức sẽ tự động định vị các vị trí từ xa thích hợp.

Hình 1.4b minh họa một tiếp cận khác, ứng dụng yêu cầu sự thực thi của một chương trình bổ trợ (auxiliary program) tại vị trí từ xa. Chương trình bổ trợ này truy xuất cơ sở dữ liệu từ xa và trả kết quả cho ứng dụng yêu cầu.

Lợi ích của cách tiếp cận thứ nhất là cung cấp sự trong suốt phân tán nhiều hơn trong khi cách tiếp cận thứ hai có thể linh động hơn nếu nhiều truy xuất cơ sở dữ liệu được yêu cầu vì ứng dụng bổ trợ có thể thực hiện tất cả các truy xuất yêu cầu và chỉ gởi kết quả về.

Một đặc tính quan trọng của hệ quản trị cơ sở dữ liệu phân tán trong hệ thống là chúng cùng loại hay khác loại. Các hệ quản trị cơ sở dữ liệu phân tán khác loại phải thêm vấn đề thông dịch giữa các mô hình dữ liệu khác nhau, các cấu trúc dữ liêu khác nhau. Đây là một vấn đề rất khó giải quyết, nên nó được khắc phục bằng cách hỗ trợ sự truyền thông giữa các thành phần truyền thông dữ liêu (data communication component DC) khác nhau. Bài toán này cũng đã được công ty Microsoft giải quyết bằng các thành phần truy xuất dữ liệu (Microsoft Data Access Components (MDAC)). Cho nên một hệ thống bao gồm các hệ quản trị cơ sở dữ liệu cục bộ khác nhau sẽ thích hợp hơn cho việc phát triển hệ thống thông tin một cách linh động và tự trị.

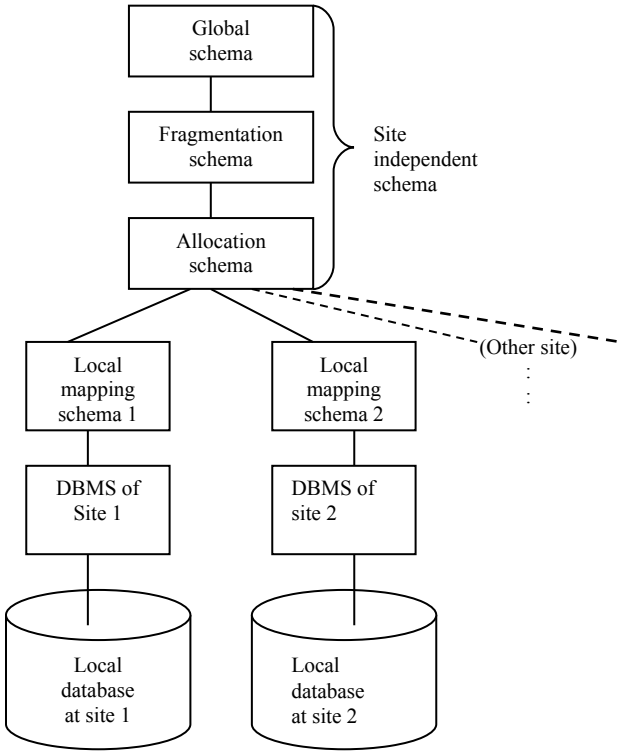
1.2 **Kiến trúc của hệ cơ sở dữ liệu phân tán và hệ quản trị cơ sở dữ liệu phân tán**

1.2.1 Kiến trúc tham khảo cho hệ cơ sở dữ liệu phân tán

Hình 1.5 mô tả kiến trúc tham khảo cho cơ sở dữ liệu phân tán. Kiến trúc tham khảo này không áp dụng cho mọi cơ sở dữ liệu phân tán. Tuy nhiên các mức của nó giúp cho ta hiểu tổ chức một cơ sở dữ liệu phân tán bất kỳ. Vì thế chúng ta sẽ phân tích và tìm hiểu tất cả các thành phần trong kiến trúc này.

Mỗi quan hệ toàn cục có thể được chia thành các thành phần không trùng nhau được gọi là các phân mảnh. Có nhiều cách để phân mảnh mà chúng ta sẽ bàn đến sau. Ánh xạ từ các quan hệ toàn cục đến các phân mảnh được định nghĩa trong lược đồ phân mảnh. Phép ánh xạ này là một-nhiều nghĩa là có một số phân mảnh tương ứng với một quan hệ toàn cục nhưng chỉ có một quan hệ toàn cục ứng với một phân mảnh. Các phân mảnh được chỉ định bởi tên quan hệ toàn cục với một chỉ mục (chỉ mục phân mảnh) ví dụ R_i chỉ phân mảnh thứ i của quan hệ toàn cục R.

Hình 1.5 Kiến trúc tham khảo cho một cơ sở dữ liệu phân tán

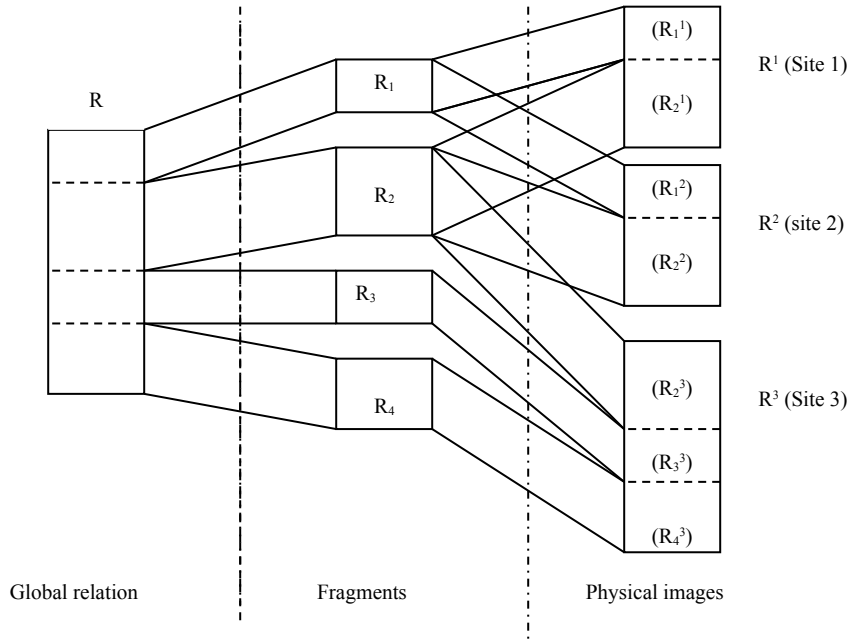


Các phân mảnh là các thành phần của các quan hệ toàn cục mà được lưu trữ vật lý tại một hay một số địa điểm. Lược đồ cấp phát (allocation scheme) xác định vị trí của một phân mảnh. Kiểu ánh xạ định nghĩa trong lược đồ định vị xác định cơ sở dữ liệu phân tán có dư thừa hay không. Trong trường hợp ánh xạ là một-nhiều thì nó dư thừa, ngược lại nếu ánh xạ có kiểu một-một thì nó không dư thừa. Tất cả các phân mảnh tương ứng với cùng một quan hệ toàn cục R và được lưu trữ tại địa điểm j tạo thành ảnh vật lý của quan hệ R tại địa điểm j. Vì thế có một ánh xạ một-một giữa một ảnh vật lý và một cặp (quan hệ toàn cục, địa điểm); các ảnh vật lý có thể được chỉ ra bởi tên quan hệ toàn cục và chỉ số địa điểm. Để phân biệt các mảnh, chúng ta sẽ sử dụng một chỉ số mũ; ví dụ, R^j chỉ ảnh vật lý của quan hệ toàn cục R tại địa điểm j.

Một ví dụ của mối quan hệ giữa các kiểu đối tượng định nghĩa ở trên được minh họa ở hình 3.2. Một lược đồ quan hệ R được phân thành bốn mảnh R_1, R_2, R_3 và R_4. Bốn phân mảnh này được lưu trữ dư thừa tại ba địa điểm trên mạng máy tính, vì thế tạo ra ba ảnh vật lý R^1, R^2, R^3.

Để làm rõ kỹ thuật này, hãy xét một bản sao của một phân mảnh tại một địa điểm và chú thích nó bằng cách sử dụng tên của quan hệ toàn cục và hai chỉ số (chỉ số phân

mảnh và chỉ số địa điểm). Ví dụ, trong hình 1.6, chú thích $R^3{}_2$ chỉ một bản sao của phân mảnh R_2 lưu trữ tại địa điểm 3.



Hình 1.6 Các phân mảnh và các ảnh vật lý đối với một quan hệ toàn cục

Cuối cùng sẽ thấy hai ảnh vật lý bất kỳ có thể được phân biệt. Trong trường hợp này ta sẽ nói một ảnh vật lý là một bản sao của một ảnh vật lý khác. Ví dụ trong hình 1.6, R^1 là một bản sao của R^2.

Kiến trúc tham khảo ở hình 1.5 đã mô tả mối quan hệ giữa các đối tượng tại ba mức trên cùng của kiến trúc này. Ba mức này độc lập vị trí, vì thế chúng không phụ thuộc vào mô hình dữ liệu của các hệ quản trị cơ sở dữ liệu cục bộ. Ở mức thấp hơn, cần ánh xạ các ảnh vật lý đến các đối tượng được thao tác bởi các hệ quản trị cơ sở dữ liệu cục bộ. Ánh xạ này được gọi là lược đồ ánh xạ cục bộ và nó phụ thuộc vào kiểu của hệ quản trị cơ sở dữ liệu cục bộ; vì thế trong hệ thống không đồng nhất, có nhiều kiểu ánh xạ cục bộ tại các vị trí khác nhau.

Kiến trúc này cung cấp một mô hình quan niệm tổng quát để hiểu được cơ sở dữ liệu phân tán. Ba đối tượng quan trọng nhất của kiến trúc này là sự tách biệt giữa sự phân mảnh dữ liệu và sự định vị dữ liệu, điều khiển dư thừa dữ liệu và tính độc lập ở các hệ quản trị cơ sở dữ liệu cục bộ.

- *Sự tách biệt quan niệm phân mảnh dữ liệu và quan niệm định vị dữ liệu*: Sự tách biệt này giúp ta phân biệt hai mức khác nhau của sự trong suốt phân tán được gọi là sự trong suốt phân tán và sự trong suốt định vị. Sự trong suốt phân tán là mức độ cao nhất của sự trong suốt và bao gồm các yếu tố mà người sử dụng và các lập trình viên làm việc trên các quan hệ toàn cục. Sự trong suốt định vị là mức thấp hơn và yêu cầu người sử dụng và các lập trình viên làm việc trên các phân mảnh thay vì trên các quan hệ toàn cục. Tuy nhiên họ không cần biết các phân mảnh này lưu trữ ở đâu. Sự tách biệt hai quan niệm phân mảnh và định vị rất phù hợp trong thiết kế cơ sở dữ liệu phân tán vì sự xác định các thành phần thích hợp của dữ liệu được nhận biết từ bài toán định vị tối ưu.

- *Điều khiển tường minh sự dư thừa dữ liệu*: Kiến trúc tham khảo cung cấp một sự điều khiển tường minh cho sự dư thừa dữ liệu tại mức phân mảnh. Ví dụ trong hình 1.6 hai ảnh vật lý R^2_2 và R^3_2 trùng lắp nghĩa là chúng chứa chung dữ liệu. Định nghĩa các phân mảnh một cách tách biệt khi xây dựng các khối vật lý cho phép tham khảo tường minh đến từng phần trùng lắp này tức phân mảnh nhân bản R_2. Điều khiển sự dư thừa dữ liệu rất hữu dụng trong một số khía cạnh quản trị cơ sở dữ liệu phân tán.

- *Tính độc lập tại các hệ quản trị cơ sở dữ liệu cục bộ*: Tính chất này gọi là sự trong suốt ánh xạ cục bộ, nó cho phép nghiên cứu một số vấn đề quản trị cơ sở dữ liệu mà không quan tâm đến mô hình dữ liệu cụ thể tại các hệ quản trị cơ sở dữ liệu cục bộ.

Một kiểu trong suốt khác liên quan chặt chẽ tới sự trong suốt định vị là sự trong suốt nhân bản. Sự trong suốt nhân bản có nghĩa là người sử dụng không nhận thấy được sự nhân bản của các phân mảnh.

1.2.2 Kiến trúc của hệ quản trị CSDL phân tán.

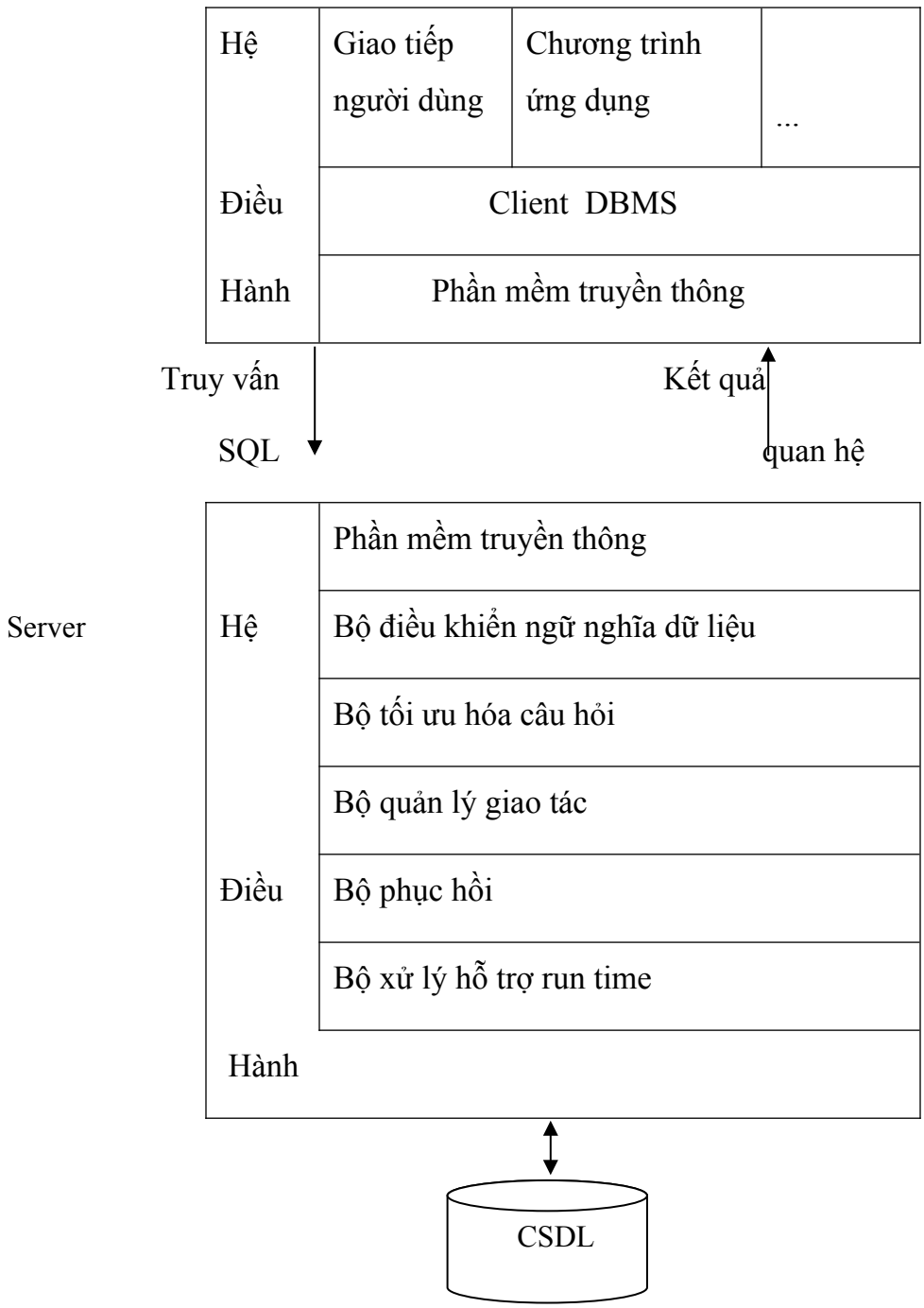
Phần này sẽ xem xét chi tiết các kiến trúc hệ thống là hệ khách/chủ (client/server), các hệ cơ sở dữ liệu phân tán và các phức hệ cơ sở dữ liệu.

1.2.2.1 Các hệ khách/chủ (Client/Server)

Trong hệ khách/chủ, ta phân biệt chức năng cần được cung cấp và chia những chức năng này thành hai lớp: chức năng chủ, chức năng khách. Nó cung cấp một kiến trúc hai mức, tạo dễ dàng cho việc quản lý mức độ phức tạp của các hệ quản trị cơ sở dữ liệu hiện đại và độ phức tạp của việc phân tán dữ liệu.

Vì thế, có thể nghiên cứu những khác biệt về chức năng khách và chức năng chủ. Điều đầu tiên phải chú ý là máy chủ thực hiện phần lớn các công việc quản lý dữ liệu. Điều này có nghĩa là mọi việc xử lý và tối ưu hóa vấn tin, quản lý giao tác và quản lý thiết bị lưu trữ đều được thực hiện tại máy chủ. Khách, ngoài giao diện và ứng dụng, sẽ có một module quản trị cơ sở dữ liệu, khách chịu trách nhiệm quản lý dữ liệu được gởi đến và đôi khi cả việc quản lý các khoá chốt giao tác.

Kiến trúc khách/chủ được biểu diễn trong hình 1.7. Kiến trúc này rất thông dụng trong các hệ thống quan hệ, ở đó việc giao tiếp giữa khách và chủ thông qua các câu lệnh SQL. Nói cách khác, khách sẽ chuyển các yêu cầu cho máy chủ mà không tìm hiểu và tối ưu hoá chúng. Máy chủ thực hiện hầu hết các công việc và trả quan hệ kết quả về cho khách.



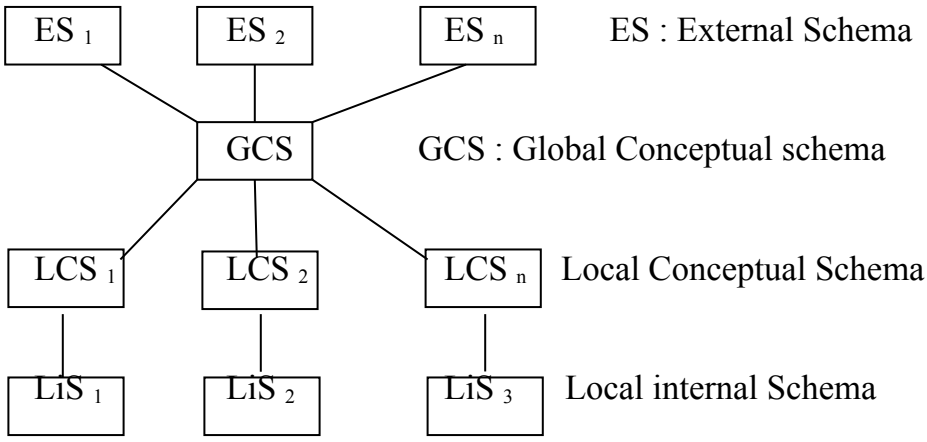
Hình 1.7 Kiến trúc khách/chủ

1.2.2.2 Hệ phân tán ngang hàng

Chúng ta bắt đầu mô tả kiến trúc này bằng cách xem xét hình ảnh tổ chức dữ liệu. Trước tiên ta chú ý rằng tổ chức dữ liệu vật lý trên mỗi máy có thể khác nhau. Vì thế cần có một định nghĩa nội tại riêng tại mỗi vị trí được gọi là lược đồ nội tại cục bộ LIS

(Local Internal Schema). Hình ảnh của mô hình dữ liệu toàn cục được mô tả bằng lược đồ quan niệm toàn cục GCS (Global Conceptual Schema). Để xử lý hiện tượng nhân bản và phân mảnh, cần phải mô tả việc tổ chức logic của dữ liệu tại mỗi vị trí, vì thế cần có một tầng thứ ba được gọi là lược đồ quan niệm cục bộ LCS (Local Conceptual Schema). Do vậy trong mô hình kiến trúc này, lược đồ quan niệm toàn cục là hợp của các quan niệm cục bộ. Cuối cùng các ứng dụng và truy xuất cơ sở dữ liệu được hỗ trợ qua lược đồ ngoài ES (External Schema). Mô hình kiến trúc này được trình bày ở hình 1.8.

Các thành phần cụ thể của một hệ quản trị cơ sở dữ liệu phân tán gồm hai thành phần chính (minh họa ở hình 1.9): bộ phận giao tiếp người sử dụng (user processor) và bộ phận xử lý dữ liệu (data processor).



Hình 1.8 Kiến trúc tham khảo cơ sở dữ liệu phân tán

Các thành phần của bộ phận giao tiếp người sử dụng gồm:

- Bộ phận giao tiếp (user interface handler): chịu trách nhiệm dịch các câu lệnh người sử dụng và định dạng dữ liệu kết quả để chuyển cho người sử dụng.

- Bộ phận kiểm soát dữ liệu ngữ nghĩa (semantic data controller): sử dụng các ràng buộc toàn vẹn và thông tin quyền hạn, được định nghĩa như thành phần của lược đồ quan niệm toàn cục để kiểm tra xem các câu truy vấn có thể xử lý được hay không.

- Bộ phận phân rã và tối ưu hoá vấn tin toàn cục (global query optimizer and decomposer): xác định như một chiến lược hoạt động nhằm giảm thiểu chi phí,

phiên dịch các câu vấn tin toàn cục thành các câu vấn tin cục bộ bằng cách sử dụng các lược đồ quan niệm toàn cục, lược đồ quan niệm cục bộ và thư mục toàn cục. Bộ phận tối ưu vấn tin toàn cục còn chịu trách nhiệm tạo ra một chiến lược thực thi tốt nhất cho phép nối phân tán.

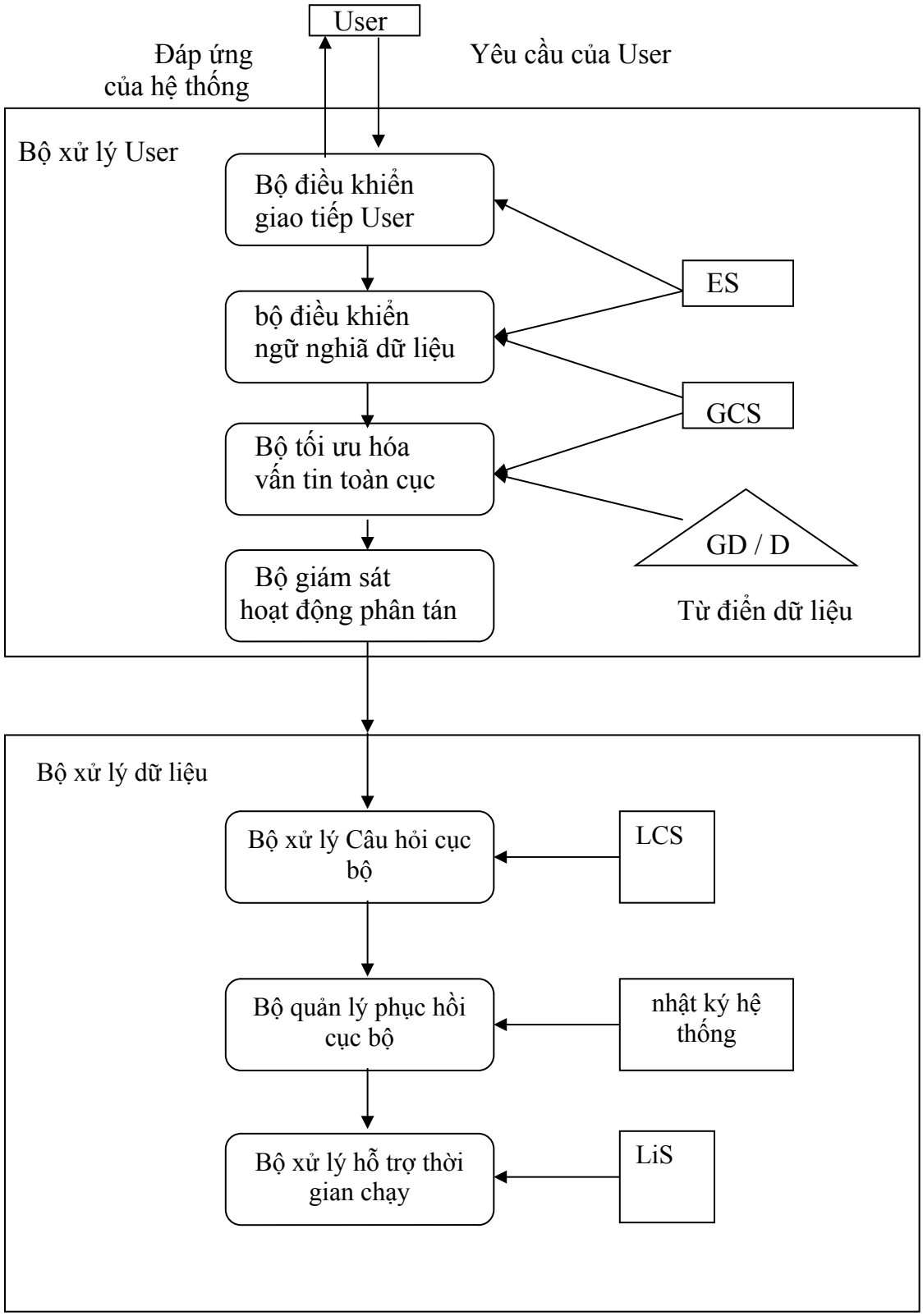
- Bộ phận giám sát hoạt động phân tán (distributed execution monitor): điều phối việc thực hiện phân tán các yêu cầu người sử dụng và cũng được gọi là bộ quản lý giao tác phân tán (distributed transaction manager).

Thành phần chủ yếu thứ hai của hệ quản trị cơ sở dữ liệu phân tán là bộ xử lý dữ liệu (data processor), bao gồm các thành phần:

- Bộ phận tối ưu hoá vấn tin cục bộ (local query optimizer): thường hoạt động như bộ chọn đường truy xuất, chịu trách nhiệm chọn ra một đường truy xuất thích hợp nhất để truy xuất các mục dữ liệu.

- Bộ phận khôi phục cục bộ (local recovery manager) bảo đảm cho các cơ sở dữ liệu cục bộ vẫn duy trì được tính nhất quán ngay cả khi có sự cố xảy ra.

- Bộ phận hỗ trợ lúc thực thi (run-time support processor): truy xuất cơ sở dữ liệu tùy thuộc vào các lệnh trong lịch biểu do bộ phận tối ưu vấn tin tạo ra. Nó chính là bộ giao tiếp với hệ điều hành và chứa bộ quản lý vùng đệm cơ sở dữ liệu, chịu trách nhiệm quản lý vùng đệm và quản lý việc truy xuất dữ liệu.



Hình 1.9 Kiến trúc của hệ quản trị cơ sở dữ liệu phân tán.