



Cơ sở dữ liệu phân tán

Phạm Minh Khan

pmkhan@hcmunre.edu.vn




Chương 1: Tổng quan về cơ sở dữ liệu phân tán

1. Sơ lược về mạng máy tính
2. Cơ sở dữ liệu phân tán là gì ?
3. Các đặc trưng của các hệ thống phân tán
4. Hệ quản trị cơ sở dữ liệu phân tán
5. Kiến trúc tham khảo cho hệ cơ sở dữ liệu phân tán
6. Lưu trữ cơ sở dữ liệu phân tán
7. Các loại truy xuất CSDL phân tán



Sơ lược về mạng máy tính

- Một mạng máy tính là một tập các máy tính tự vận hành, được kết nối lại và có khả năng trao đổi thông tin với nhau.
 - Các máy tính trên một mạng thường được gọi là các nút hay các trạm (site) được kết nối lại với nhau thông qua đường truyền.
- 

Sơ lược về mạng máy tính

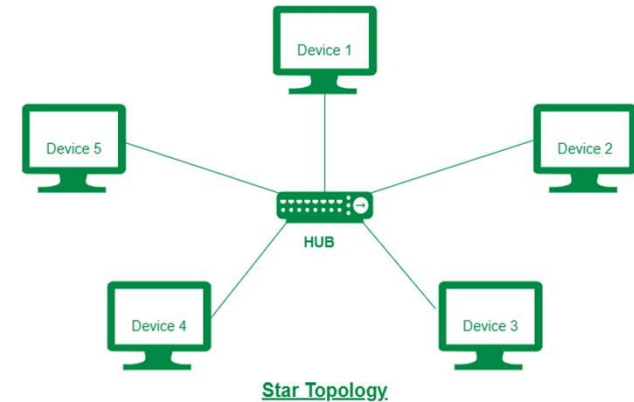
➤ Các loại mạng máy tính cơ bản:

✓ Mạng hình sao (Star Topology)

❖ Mạng hình sao là một mô hình mạng bao gồm một thiết bị làm trung tâm và các nút thông tin chịu sự điều khiển của trung tâm đó.

❖ Ưu điểm:

- ✓ Khi có lỗi xảy ra ở một máy trạm nào đó thì cả hệ thống vẫn hoạt động bình thường.
- ✓ Tốc độ mạng hình sao khá nhanh.
- ✓ Cấu trúc mạng khá đơn giản giúp dễ dàng kiểm tra, sửa chữa khi gặp sự cố
- ✓ Mạng này có thể thu hẹp hoặc mở rộng khi cần
- ✓ Giúp hạn chế được các yếu tố gây ngưng trệ mạng vì kiểu liên kết này cho phép nối trực tiếp các máy tính với Hub



Nguồn: www.geeksforgeeks.org

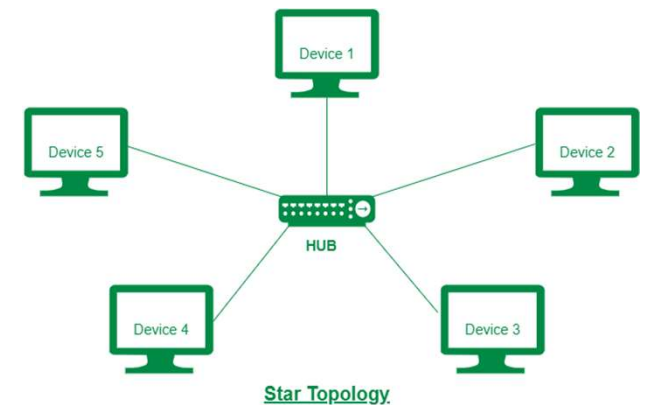
Sơ lược về mạng máy tính

➤ Các loại mạng máy tính cơ bản:

✓ **Mạng hình sao (Star Topology)**

❖ Nhược điểm:

- ✓ Thiết bị trung tâm là yếu tố chủ chốt, vì vậy khi nó bị sự cố thì tất cả các thiết bị đều chịu ảnh hưởng.
- ✓ Khoảng cách kết nối hạn chế (chỉ khoảng 100 mét)
- ✓ Tồn chi phí dây mạng và thiết bị trung gian.



Nguồn: www.geeksforgeeks.org

Sơ lược về mạng máy tính

➤ Mạng dạng tuyến (Bus Topology)

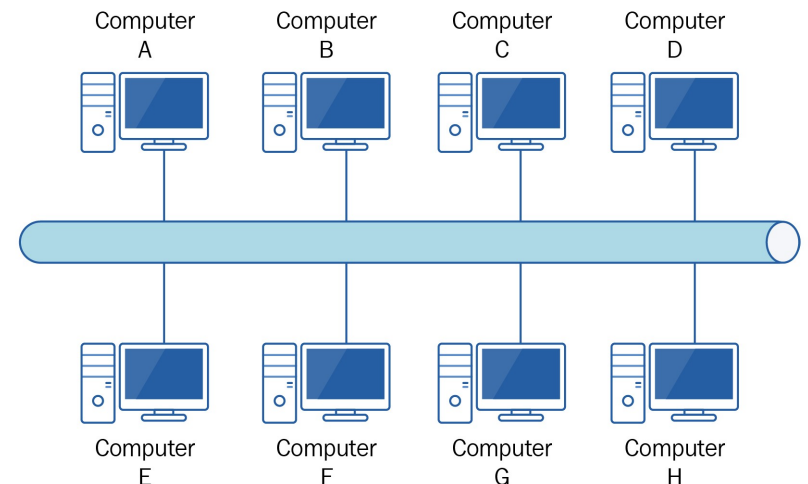
❖ Tất cả các thiết bị như máy chủ, máy trạm, các nút thông tin đều được liên kết với nhau trên một đường dây cáp chính để truyền dữ liệu.

❖ Ưu điểm:

- ✓ Dễ dàng lắp đặt.
- ✓ Không bị giới hạn về độ dài dây cáp.

❖ Nhược điểm:

- ✓ Rất khó để xác định nơi xảy ra lỗi
- ✓ Với lưu lượng lớn, dễ dẫn đến tình trạng tắc nghẽn trên đường truyền.

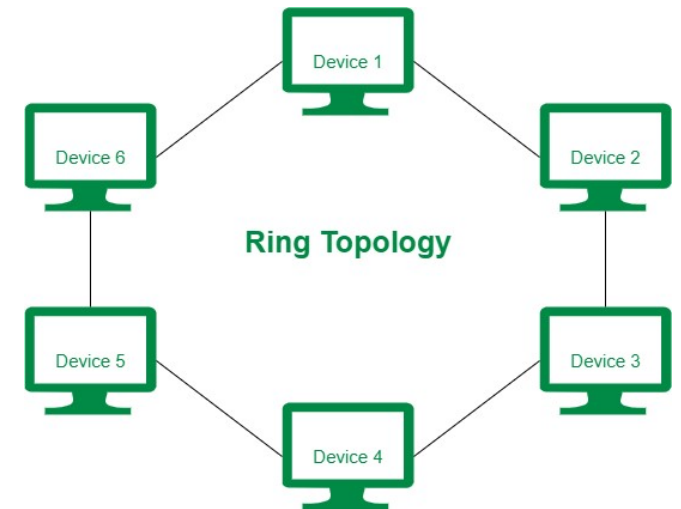


Nguồn: subscription.packtpub.com

Sơ lược về mạng máy tính

➤ Mạng dạng vòng (Ring Topology)

- ❖ Nơi các thiết bị được kết nối thành một vòng tròn khép kín thông qua dây cáp.
- ❖ Ưu điểm:
 - ✓ Dễ dàng mở rộng hệ thống LAN ra xa hơn.
 - ✓ Tiết kiệm được chiều dài dây cáp (cable)
 - ✓ Tốc độ mạng nhanh hơn mạng dạng tuyến (Bus Topology)
- ❖ Nhược điểm:
 - ✓ Do kết nối khép kín nên 1 thiết bị trục trặc là hệ thống ngừng hoạt động
 - ✓ Khó kiểm tra để tìm lỗi khi có sự cố.

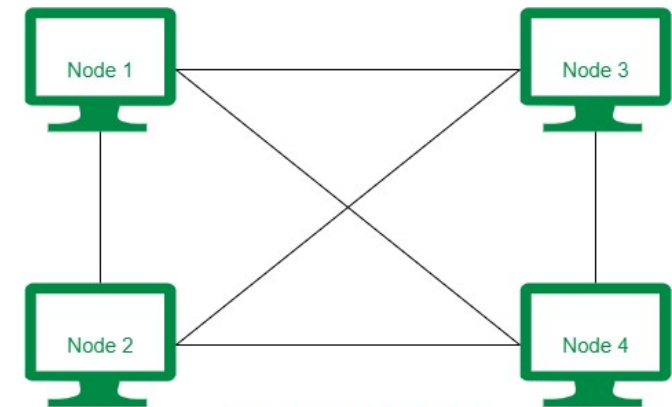


Nguồn: www.geeksforgeeks.org

Sơ lược về mạng máy tính

➤ Mạng dạng lưới (Mesh Topology)

- ❖ Mỗi một máy tính sẽ được liên kết với tất cả các máy còn lại trên hệ thống mà không cần phải nối qua Hub hay Switch.
- ❖ Ưu điểm:
 - ✓ Các máy tính trong hệ thống này hoạt động độc lập
 - ✓ Nó tương tự như mạng hình sao nhưng được mở rộng với phạm vi lớn hơn
- ❖ Nhược điểm:
 - ✓ Việc quản lý hệ thống mạng sẽ khá phức tạp
 - ✓ Gây tốn tài nguyên về bộ nhớ và về việc xử lý của các máy trạm trong hệ thống.

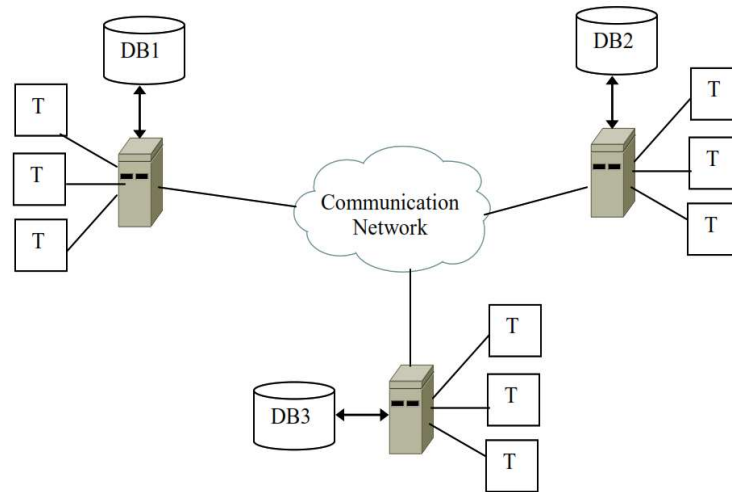


Mesh Topology.

Nguồn: www.geeksforgeeks.org

Cơ sở dữ liệu phân tán là gì ?

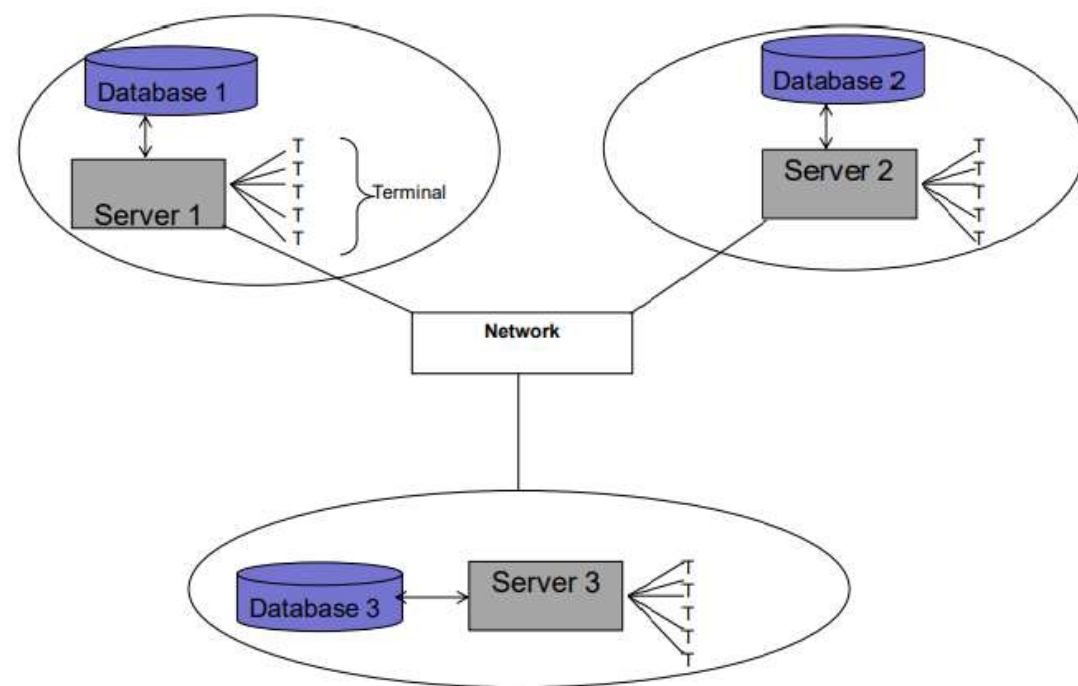
- ❖ Một cơ sở dữ liệu phân tán là một **tập hợp dữ liệu của một hệ thống** nhưng **được phân bố trên nhiều địa điểm (site)** của một mạng máy tính.
 - ✓ Sự phân tán: dữ liệu không lưu trữ trên cùng một địa điểm vì thế chúng ta có thể phân biệt nó với cơ sở dữ liệu tập trung.
 - ✓ Mối tương quan luận lý (logical correlation): 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



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

Ví dụ 1:

- Một ngân hàng có ba chi nhánh đặt tại các vị trí khác nhau.
- Tại mỗi chi nhánh (site) có một máy tính điều khiển một số máy giao dịch đầu cuối (teller terminal) và cơ sở dữ liệu của chi nhánh đó.
- Tại mỗi site được đặt một phần cơ sở dữ liệu phân tán.
- Các máy tính được nối với nhau bởi một mạng truyền thông.
- Các nút trong một mạng phân tán có hai chức năng:
 - ✓ Xử lý thông tin tại vị trí mà nó quản lý
 - ✓ Tham gia vào việc xử lý các yêu cầu về thông tin cần truy cập qua nhiều địa điểm. Chẳng hạn, việc lên danh sách tất cả nhân viên của ngân hàng. Yêu cầu này đòi hỏi tất cả các máy tính ở các chi nhánh của công ty đều phải hoạt động để cung cấp thông tin.

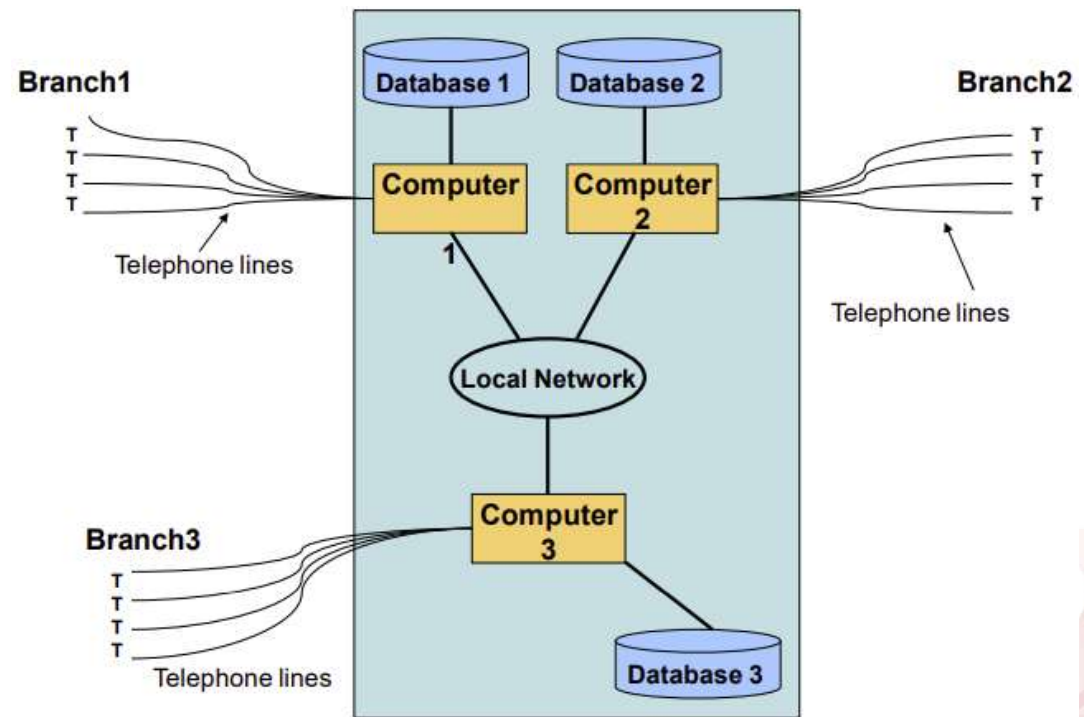


CSDL phân tán trên một mạng phân tán về địa lý

Ví dụ 2:

Xét một ngân hàng như ở ví dụ 1. Nhưng:

- Các máy tính với CSDL của chúng ở mỗi chi nhánh được chuyển đến cùng một tòa nhà.
- Các máy tính này được kết nối với nhau bởi một mạng cục bộ với băng thông rộng.
- Các máy giao dịch đầu cuối được kết nối với máy tính tương ứng của chúng qua đường dây điện thoại
- Mỗi máy tính và CSDL của nó tạo nên một site của mạng cục bộ.

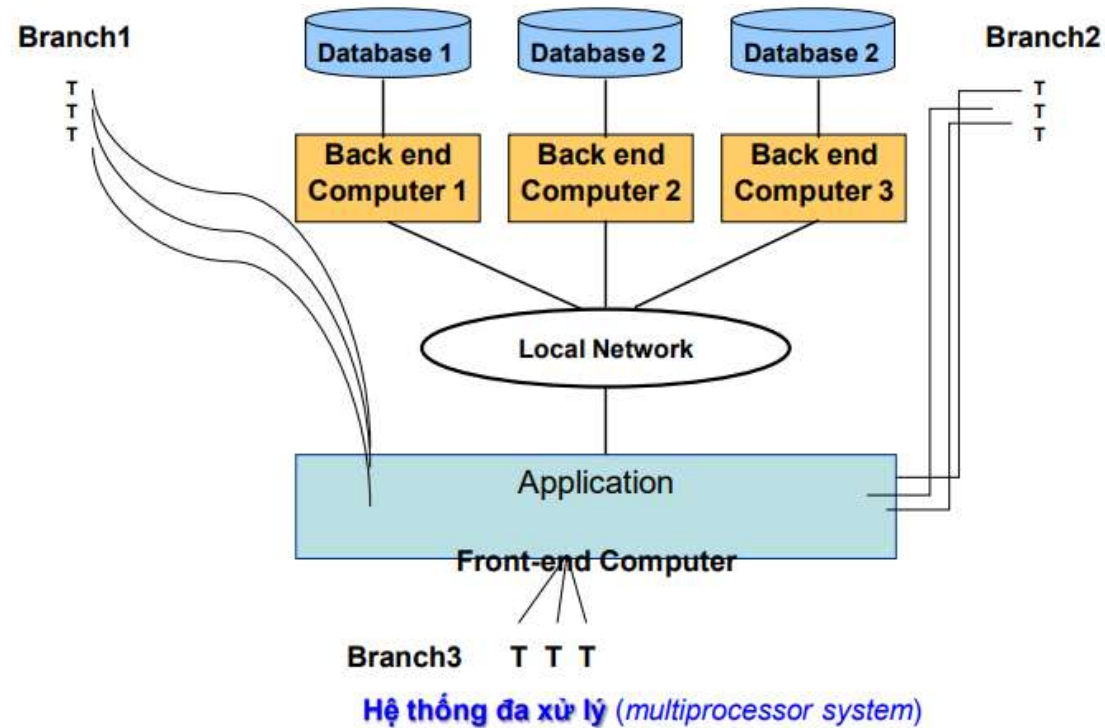


CSDL phân tán trên một mạng cục bộ

Ví dụ 3:

Xét hệ thống ngân hàng như trên. Nhưng:

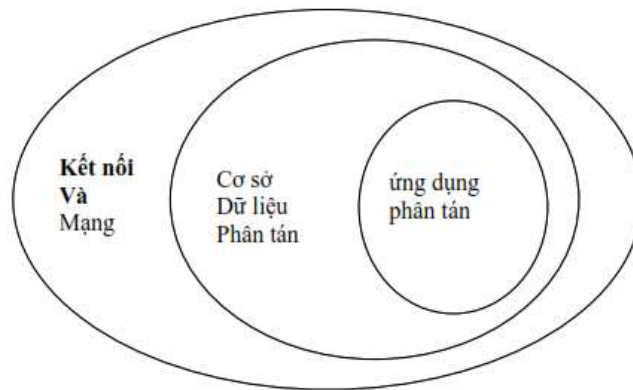
- Dữ liệu của các chi nhánh được phân tán trên 3 back-end computers, các máy tính này sẽ thực hiện chức năng quản trị CSDL
- Các chương trình ứng dụng được thực hiện bởi một máy tính khác (front-end computer), máy tính này đóng vai trò giao diện và yêu cầu back-end computers thực hiện các lệnh của người sử dụng.



Cơ sở dữ liệu phân tán là gì ?

❖ Định nghĩa:

- 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 trải 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.



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

Các đặc trưng của các hệ thống phân tán

❖ So sánh các tính chất đặc trưng của CSDL tập trung và CSDL phân tán

| Tính chất đặc trưng | Cơ sở dữ liệu tập trung | Cơ sở dữ liệu phân tán |
|----------------------|---|--|
| Điều khiển tập trung | <ul style="list-style-type: none">- Khả năng cung cấp sự điều khiển tập trung trên các tài nguyên thông tin.- Cần có người quản trị cơ sở dữ liệu | <ul style="list-style-type: none">- Cấu trúc điều khiển phân cấp: quản trị CSDL toàn cục và quản trị CSDL cục bộ phân tán |
| Độc lập dữ liệu | <ul style="list-style-type: none">- Tổ chức dữ liệu trong suốt với các lập trình viên.- 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 | <p>Giảm thiểu sự dư thừa dữ liệu do:</p> <ul style="list-style-type: none">- Tính nhất quán dữ liệu cao- Tiết kiệm dung lượng nhớ. | <ul style="list-style-type: none">- 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 đặc trưng của các hệ thống phân tán

❖ So sánh các tính chất đặc trưng của CSDL tập trung và CSDL phân tán (tt)

| Tính chất đặc trưng | Cơ sở dữ liệu tập trung | Cơ sở dữ liệu phân tán |
|--|---|--|
| 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. |

Các đặc trưng của các hệ thống phân tán

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



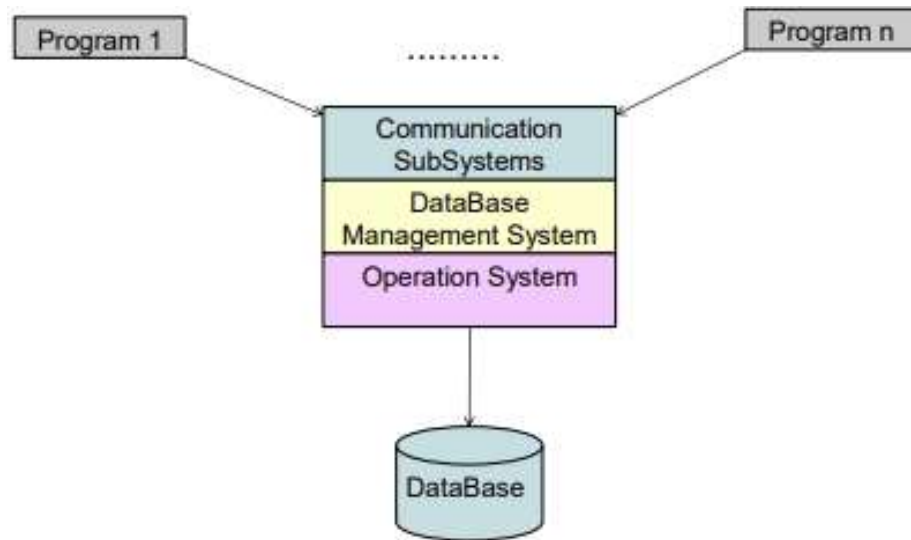
- ✓ Lý do tổ chức kinh tế
- ✓ Lý do kết nối các cơ sở dữ liệu hiện có
- ✓ Lý do tăng trưởng tổ chức
- ✓ Lý do tải truyền thông
- ✓ Đánh giá về hiệu suất
- ✓ Độ tin cậy và tính hiệu quả

Các đặc trưng của các hệ thống phân tán

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

| Ưu điểm | Nhược điểm |
|--|-----------------------------|
| Chia sẻ dữ liệu và điều khiển phân tán | Chi phí phát triển phần mềm |
| Độ tin cậy và tính sẵn sàng | Khó phát hiện lỗi |
| Tăng tốc độ xử lý truy vấn | Chi phí xử lý tăng |

Hệ quản trị cơ sở dữ liệu tập trung



Kiến trúc tổng quát của một hệ quản trị CSDL tập trung

Tầng giao diện

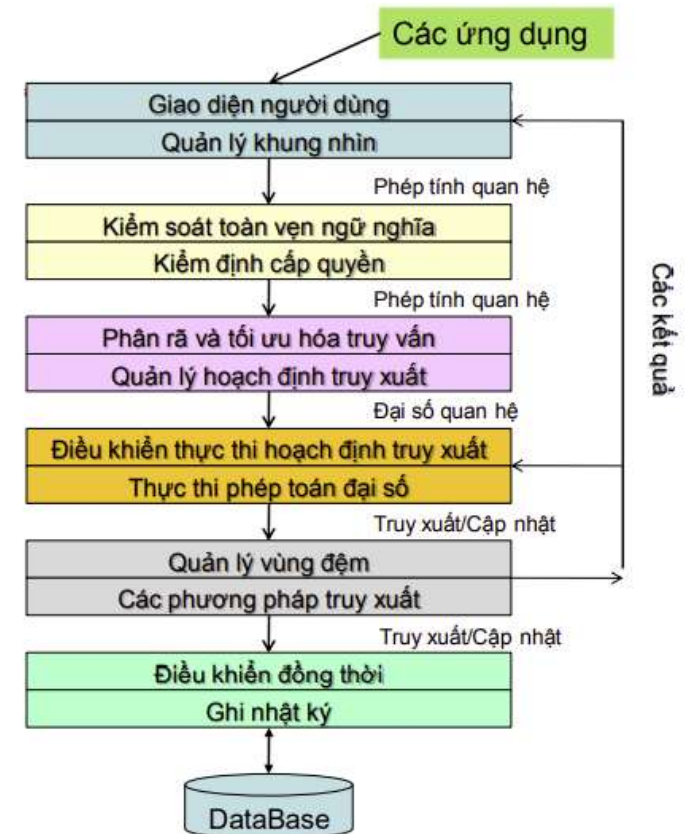
Tầng điều khiển

Tầng biên dịch

Tầng thực thi

Tầng truy xuất dữ liệu

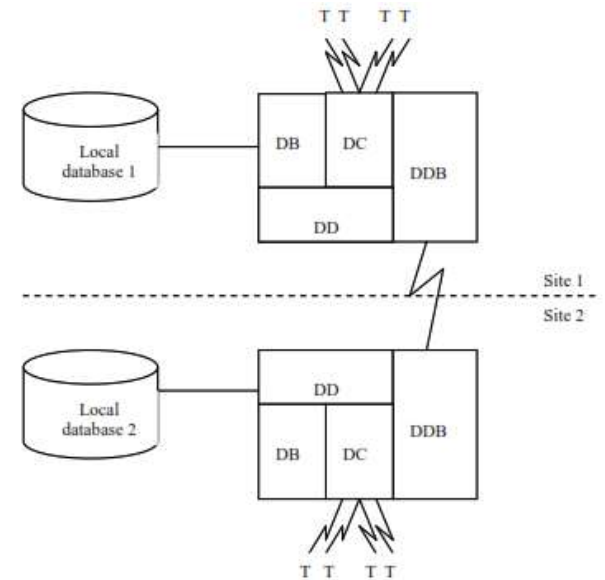
Tầng duy trì nhất quán



Các tầng chức năng của một Hệ QTCSDL quan hệ

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 thành phần cơ bản cho việc xây dựng một CSDL phân tán là
 - ✓ Thành phần quản trị cơ sở dữ liệu (DB Database Management)
 - ✓ Thành phần truyền dữ liệu (DC Data Communication)
 - ✓ Từ điển dữ liệu (DD Data Dictionary) mở rộng để biểu diễn thông tin về sự phân tán dữ liệu trên mạng.
 - ✓ Thành phần cơ sở dữ liệu phân tán (DDB Distributed Database)

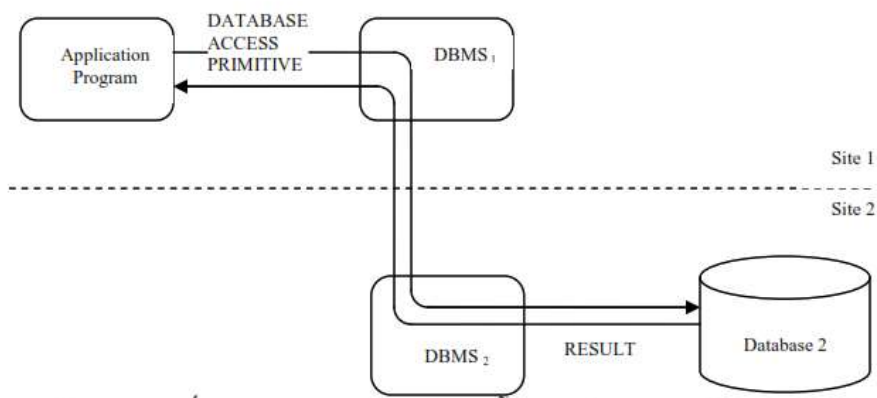


Hệ quản trị cơ sở dữ liệu phân tán

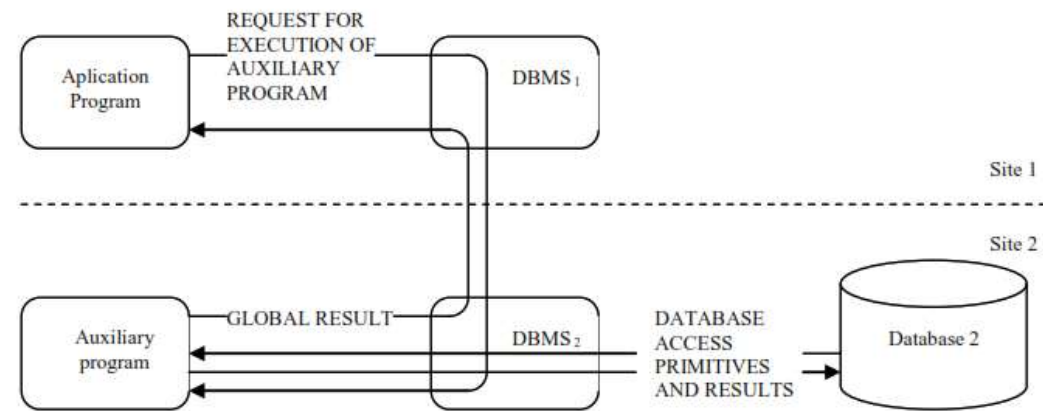
➤ Các chức năng cơ bản của cơ sở dữ liệu phân tán

- Dịch vụ truy xuất cơ sở dữ liệu từ xa
- Mức độ trong suốt của sự phân tán
- Hỗ trợ cho việc quản trị và điều khiển CSDL phân tán
- Hỗ trợ cho việc điều khiển đồng thời và phục hồi các giao tác phân tán

Các loại truy xuất CSDL phân tán

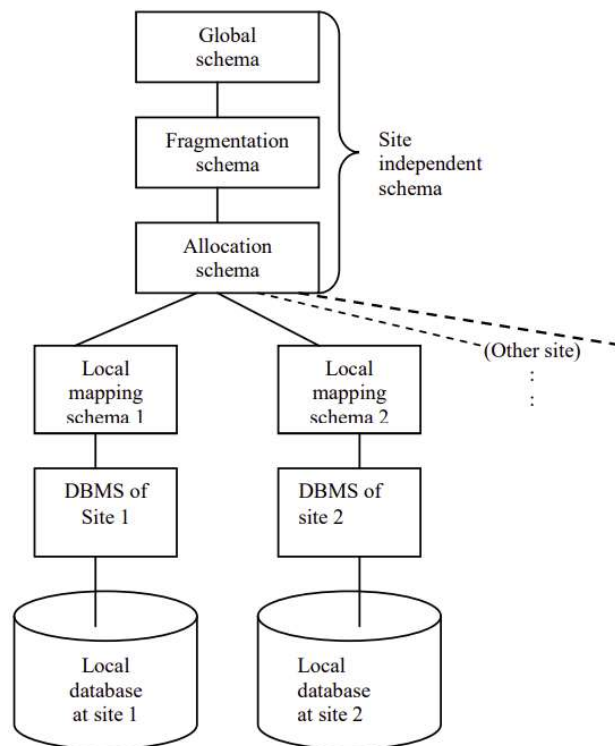


Cách 1: Truy xuất từ xa qua các lệnh có sẵn của hệ quản trị cơ sở dữ liệu



Cách 2: Truy xuất từ xa qua chương trình hỗ trợ

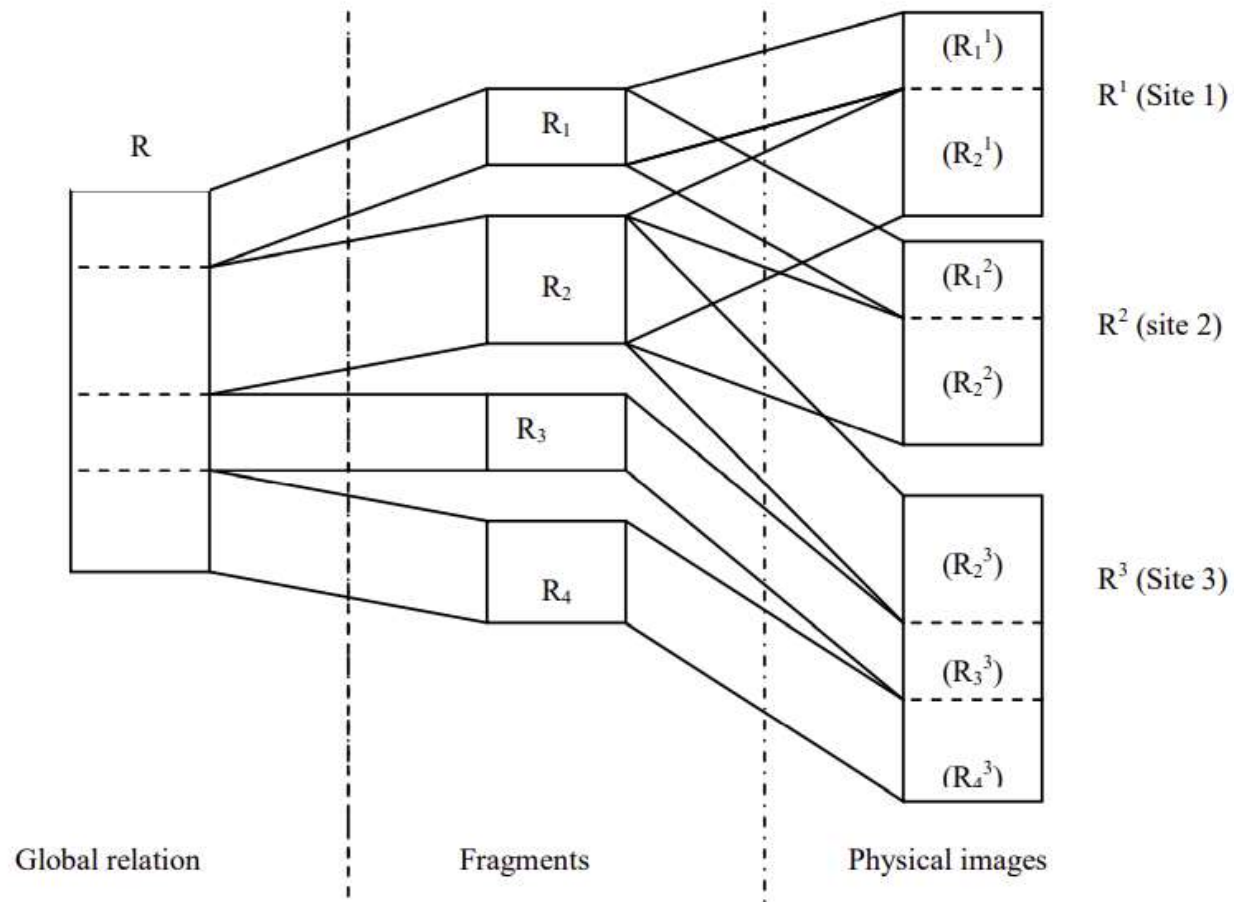
Kiến trúc tham khảo cho hệ cơ sở dữ liệu phân tán



Kiến trúc tham khảo cho một cơ sở dữ liệu

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

Các phân mảnh và các ảnh vật lý



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

Kiến trúc của hệ quản trị CSDL phân tán

❖ Các hệ khách/chủ (Client/Server)

Trong hệ khách/chủ: chia những chức năng này thành hai lớp: chức năng chủ, chức năng khách

Client

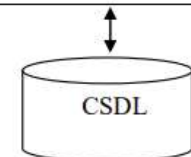
| | | | |
|------|-----------------------|-----------------------|-----|
| Hệ | Giao tiếp người dùng | Chương trình ứng dụng | ... |
| Điều | Client DBMS | | |
| Hành | Phần mềm truyền thông | | |

Truy vấn
SQL

Kết quả
quan hệ

Server

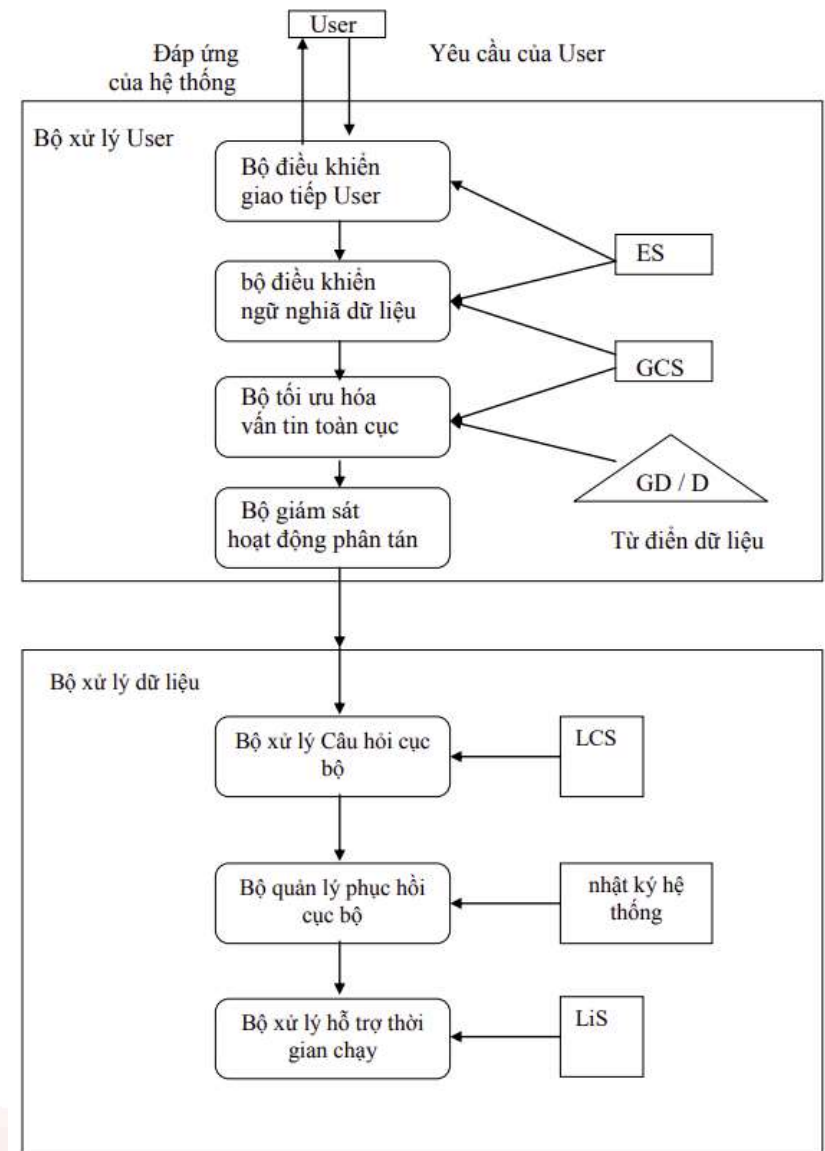
| | |
|------|---------------------------------|
| Hệ | Phần mềm truyền thông |
| | Bộ điều khiển ngữ nghĩa dữ liệu |
| | Bộ tối ưu hóa câu hỏi |
| | Bộ quản lý giao tác |
| Điều | Bộ phục hồi |
| | Bộ xử lý hỗ trợ run time |
| Hành | |



Kiến trúc của hệ quản trị CSDL phân tán

❖ Kiến trúc của hệ quản trị cơ sở dữ liệu phân tán

Xử lý trên phân tán



Lưu trữ cơ sở dữ liệu phân tán

Lưu trữ cơ sở dữ liệu phân tán được quản lý theo hai cách:

- Nhân bản (Replication)
- Phân mảnh (Fragmentation)

Lưu trữ cơ sở dữ liệu phân tán

➤ Nhân bản (Replication)

- ✓ Trong cơ sở dữ liệu nhân bản, hệ thống lưu trữ các bản sao của dữ liệu trên các site khác nhau. Nếu toàn bộ cơ sở dữ liệu có sẵn trên nhiều trang web, nó là một cơ sở dữ liệu hoàn toàn dư thừa.
 - ✓ Ưu điểm của cơ sở dữ liệu nhân bản là nó làm tăng tính khả dụng của dữ liệu trên các site khác nhau và cho phép xử lý các yêu cầu truy vấn song song.
 - ✓ Tuy nhiên, cơ sở dữ liệu nhân bản có nghĩa là dữ liệu yêu cầu cập nhật liên tục và đồng bộ hóa với các site khác để duy trì bản sao cơ sở dữ liệu chính xác. Bất kỳ thay đổi nào được thực hiện trên một site phải được ghi lại trên các site khác, nếu không sẽ xảy ra sự mâu thuẫn.
- => Cập nhật liên tục gây ra nhiều chi phí máy chủ và làm phức tạp việc kiểm soát đồng thời, vì phải kiểm tra nhiều truy vấn đồng thời trong tất cả các trang có sẵn.

Lưu trữ cơ sở dữ liệu phân tán

➤ Phân mảnh (Fragmentation)

- ✓ Sự phân mảnh trong CSDL phân tán có nghĩa là chúng được chia thành các phần nhỏ hơn và được lưu ở các site khác nhau.
- ✓ Điều kiện tiên quyết để phân mảnh là đảm bảo rằng **các mảnh sau này có thể được tái tạo lại thành quan hệ ban đầu** mà không làm mất dữ liệu.
- ✓ Ưu điểm của phân mảnh là không có bản sao dữ liệu, điều này giúp cho dữ liệu luôn được nhất quán.
- ✓ Có 3 loại phân mảnh:
 - Phân mảnh ngang - Lược đồ quan hệ được phân mảnh thành các nhóm hàng và mỗi nhóm (tuple) được gán cho một phân mảnh.
 - Phân mảnh dọc - Lược đồ quan hệ được phân mảnh thành các lược đồ nhỏ hơn và mỗi đoạn chứa một khóa ứng viên chung để đảm bảo một phép nối không mất dữ liệu.
 - Phân mảnh hỗn hợp: là kết hợp cả phân mảnh ngang và phân mảnh dọc

Hệ quản trị cơ sở dữ liệu phân tán

Ví dụ:

Các kiểu phân mảnh– dọc

PROJ₁: thông tin về kinh phí của đề án

PROJ₂: thông tin về tên và vị trí của đề án

PROJ

| PNO | PNAME | BUDGET | LOC |
|-----|-------------------|--------|----------|
| P1 | Instrumentation | 150000 | Montreal |
| P2 | Database Develop. | 135000 | New York |
| P3 | CAD/CAM | 250000 | New York |
| P4 | Maintenance | 310000 | Paris |
| P5 | CAD/CAM | 500000 | Boston |

PROJ₁

| PNO | BUDGET |
|-----|--------|
| P1 | 150000 |
| P2 | 135000 |
| P3 | 250000 |
| P4 | 310000 |
| P5 | 500000 |

PROJ₂

| PNO | PNAME | LOC |
|-----|-------------------|----------|
| P1 | Instrumentation | Montreal |
| P2 | Database Develop. | New York |
| P3 | CAD/CAM | New York |
| P4 | Maintenance | Paris |
| P5 | CAD/CAM | Boston |

PHF – Ví dụ: phân mảnh ngang cho quan hệ PAY

PAY₁

| TITLE | SAL |
|------------|-------|
| Mech. Eng. | 27000 |
| Programmer | 24000 |

PAY₂

| TITLE | SAL |
|-------------|-------|
| Elect. Eng. | 40000 |
| Syst. Anal. | 34000 |

Tài liệu tham khảo

- Tài liệu giảng dạy cơ sở dữ liệu phân tán của PIIT
- M.TAMER OZSU and PATRICK VALDURIEZ, Principles of Distributed Database Systems, Springer, 2020
- Cơ sở dữ liệu phân tán, PGS.TS Nguyễn Mậu Hân



Thank you for listening