Chương 1

Giới thiệu

Phạm Thị Ngọc Diễm
Bộ môn HTTT - Khoa CNTT&TT
ptndiem@cit.ctu.edu.vn

Nội dung

- Tại sao phải phân tán dữ liệu
- Ví dụ CSDL phân tán
- Định nghĩa
- Ưu và nhược điểm CSDL phân tán
- Tiếp cận thiết kế CSDL phân tán
- Kiến trúc CSDL phân tán
- Tự điển dữ liệu toàn cục
- Sự trong suốt

Tại sao cần phân tán dữ liệu

- **Nhu cầu phân tán dữ liệu**
  - Cần phân tán thông tin (trong trường hợp các công ty đa quốc gia)
  - Sự gia tăng nhanh chóng thông tin (ví dụ hơn 14 lần 1990-2000),
  - Số lượng giao dịch tăng nhanh (gấp hơn 10 lần trong vòng 5 năm).

    => Cần các máy chủ CSDL có khả năng đáp ứng nhanh trên các dữ liệu lớn

Facebook generates 4 petabytes of data per day
(https://kinsta.com/blog/facebook-statistics/, 2018)

petabyte (PB) = 1.024 TB = 1 triệu GB

1,5 PB = 10 tỷ ảnh trên Facebook

20 PB = Lượng dữ liệu được Google xử lý hàng ngày trong năm 2008.

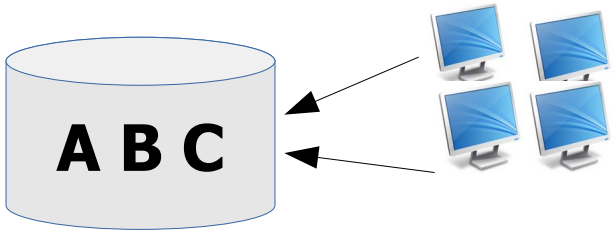
Tại sao cần phân tán dữ liệu

- **Để cải thiện tốc độ vào/ra:**
  - Phân bố dữ liệu,
  - Truy xuất song song dữ liệu,
  - Sử dụng nhiều nút (với một chi phí / hiệu suất tốt) và cho phép chúng giao tiếp qua mạng.

    => Các CSDL phân tán được phát triển nhờ vào sự tiến bộ của công nghệ hạ tầng mạng và máy trạm.

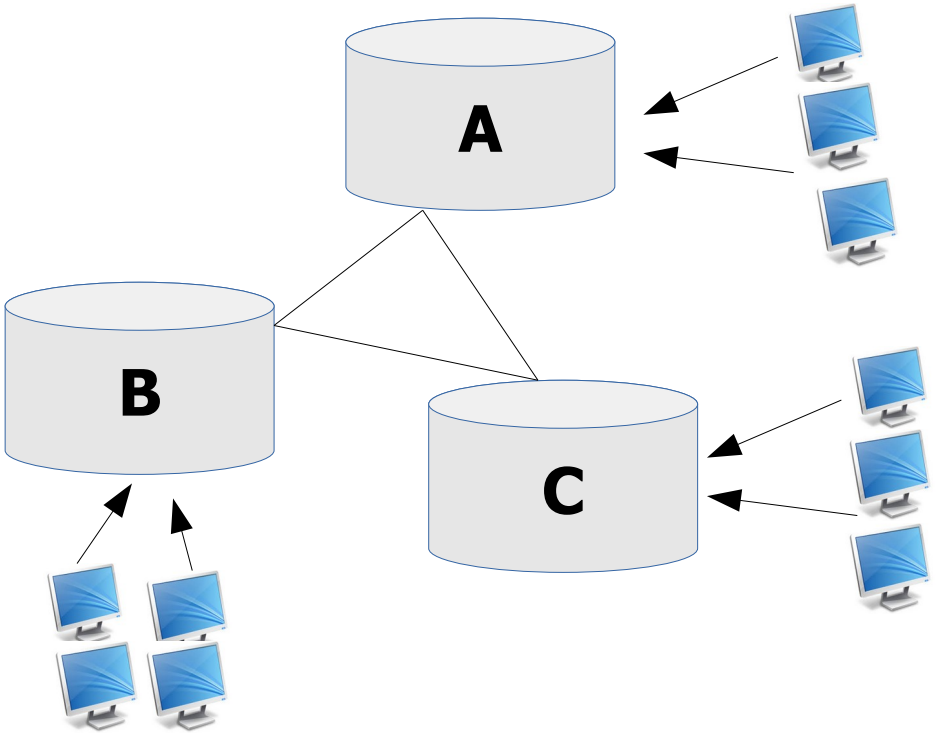
DB và D-DB

- CSDL tập trung



+ Quyền truy xuất như nhau
+ Dễ quản lý
- Cạnh tranh CSDL

- CSDL phân tán



+ Truy xuất nhanh các dữ liệu cục bộ
+ Tự trị cục bộ tại mỗi nút,
+ Quyền truy xuất như nhau, dễ quản lý
+ Có thể truy cập vào các nút khác
- Quản lý toàn cục các CSDL

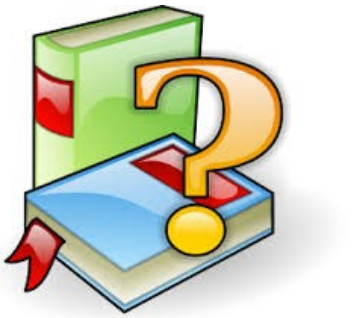
Định nghĩa

- **CSDL phân tán** (D-DB Distributed Database):

  Một CSDL phân tán là một tập các dữ liệu có quan hệ logic với nhau (gồm một lược đồ toàn cục) và được phân bố trên nhiều nút/vị trí trong một mạng máy tính

- **Hệ quản trị CSDL phân tán:**
  Hệ thống cho phép quản lý, tạo, truy xuất và thao các cơ sở dữ liệu phân tán giống như một hệ quản trị CSDL tập trung, ngoài ra nó còn cung các thành phần hỗ trợ sao cho sự phân tán là trong suốt đối với người sử dụng
  - *Mỗi nút tham gia vào việc thực thi của ít nhất một ứng dụng toàn cục thông qua mạng*
  - *Mỗi nút có khả năng xử lý tự trị cục bộ*

Định nghĩa



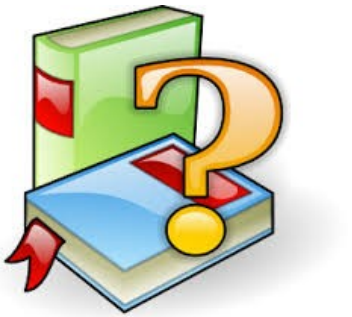
- **Nút (Site)** : Thuật ngữ site đại diện cho một vị trí luận lý trong một sơ đồ kiến trúc hoặc sơ đồ triển khai. Cụ thể, đó có thể là một máy thực sự, nhưng nó không hoàn toán đúng.

Ví dụ: chúng ta có thể triển khai bằng cách sử dụng hai site (Site-A và Site-B). Hai site đó có thể là:

- Bất kỳ hai máy nào miễn là tất cả các yêu cầu cần thiết cho các máy và các kết nối mạng của chúng được thỏa mãn.
- Trong một số trường hợp nhất định, chúng ta cũng có thể triển khai tất cả các hệ thống con đặt tại Site-A và Site-B trên cùng một máy.

=> Lưu ý: D-DB được triển khai trên một máy đơn thì vẫn là D-DB.

Định nghĩa



- ***Phân đoạn (fragmentation)*** :

  Phân đoạn là quá trình phân rã của một cơ sở dữ liệu thành một tập hợp các CSDL con. Sự phân rã này phải là sự phân rã không mất thông tin.

- **Nhân bản (replication)**:

  Nhân bản cơ sở dữ liệu là quá trình sao chép dữ liệu từ một CSDL trên một máy tính hoặc máy chủ thành một cơ sở dữ liệu trên một máy khác để tất cả người dùng chia sẻ cùng mức thông tin.

Chú ý

Không nên nhầm lẫn một CSDL phân tán với:

- Một hệ thống trong đó các cơ sở dữ liệu có thể được truy cập từ xa
- Một hệ thống multiple DB hoặc Federated DB.
  - Trong một multiple DB (M-DB), một số DB tương tác với ứng dụng thông qua một ngôn ngữ chung và không có mô hình chung.
  - Trong một Federated DB (F-DB), nhiều DB không đồng nhất được truy cập như chỉ một nhờ vào một view chung.

Ưu điểm và nhược điểm

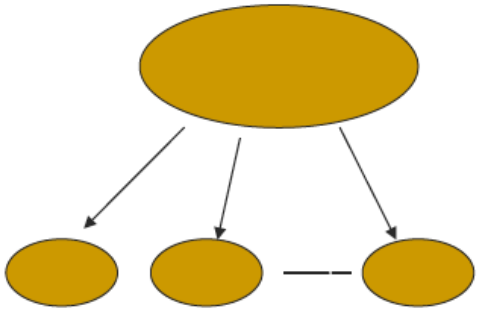
- **Ưu điểm:**
  - Cải thiện hiệu suất
  - Khả năng mở rộng hệ thống (bổ sung các nút mới)
    - Nâng cấp hệ thống (scale-up)
    - Mở rộng quy mô (scale-out)
  - Tính trong suốt (transparency)
    - Lưu trữ dữ liệu
    - Thực thi câu truy vấn
  - Sự tự trị cục bộ (tại mỗi nút)
  - Năng cao tính sẵn dùng, độ tin cậy

Ưu điểm và nhược điểm

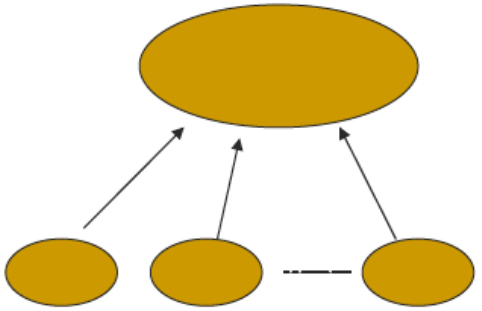
- **Hạn chế:**
  - Thiết kế và cài đặt CDSL phức tạp
  - Tính nhất quán của dữ liệu khó điều khiển
  - Vấn đề bảo mật
  - Vấn đề phát hiện và phục hồi lỗi

Tiếp cận thiết kế D-DB

- Hai tiếp cận thiết kế D-DB phổ biến:
  - tiếp cận từ trên xuống (top down)



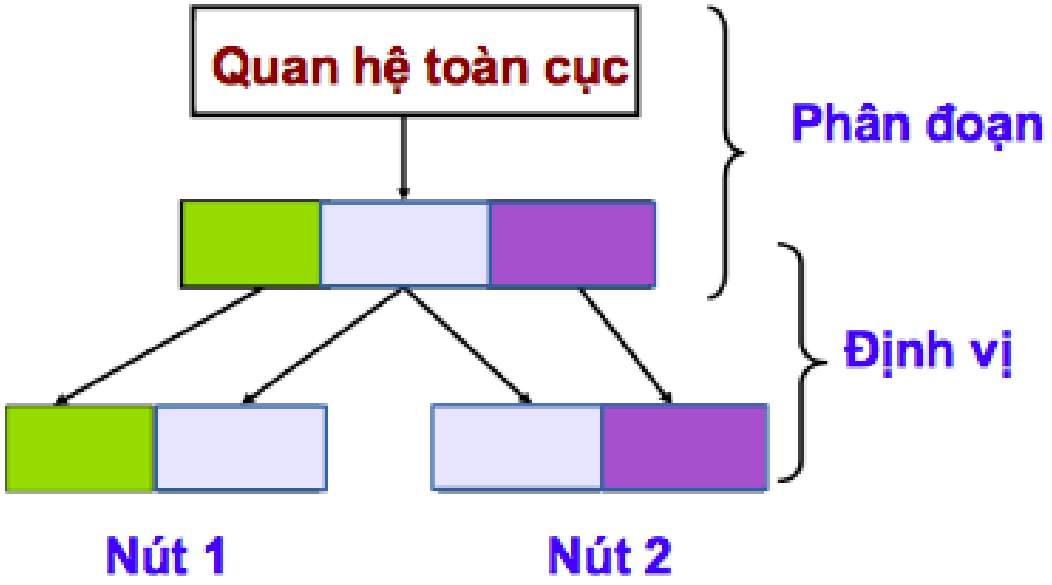
  - tiếp cận từ dưới lên (bottom up)



Tiếp cận từ trên xuống

- **Bước 1**: Xác định một lược đồ (schema) quan niệm tổng thể (global schema) của D-DB
- **Bước 2**: Phân tán trên các nút khác nhau các lược đồ quan niệm cục bộ. Sự phân bố như vậy được thực hiện theo hai bước:
  - Bước đầu tiên được gọi là phân đoạn (fragmentation)
  - Bước tiếp theo là định vị (allocation) các đoạn này trên các nút khác nhau.

Tiếp cận từ trên xuống



=> được sử dụng khi chúng ta bắt đầu từ đầu, nghĩa là chưa có CSDL nào tồn tại. Nếu có các CSDL tồn tại, *tiếp cận từ dưới lên được sử dụng.*

Tiếp cận từ dưới lên

- Cách tiếp cận này dựa trên thực tế là sự phân tán đã được thực hiện,

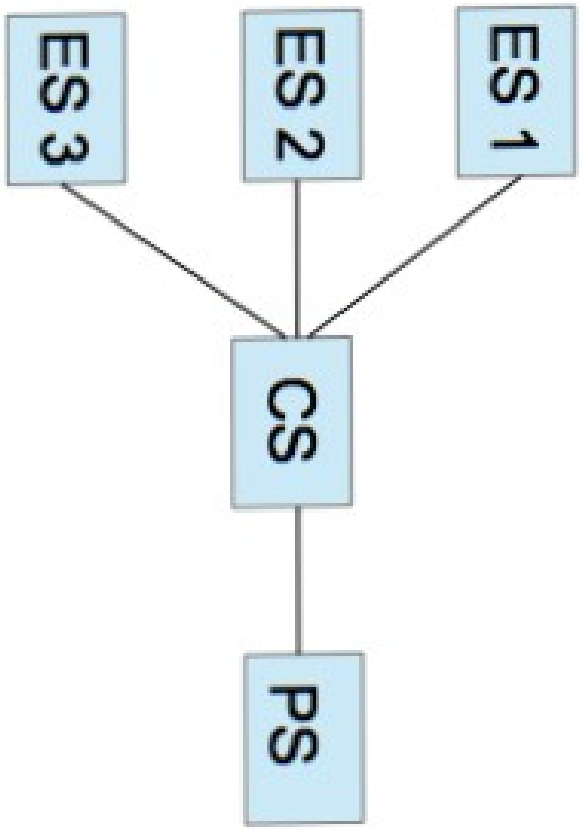
  => Cần tích hợp các CSDL khác nhau vào một cơ sở dữ liệu toàn cục.

  - Lược đồ quan niệm cục bộ đã tồn tại
  - Liên kết chúng trong một sơ đồ quan niệm tổng thể.

=> Được dùng cho federated database

Kiến trúc D - DB

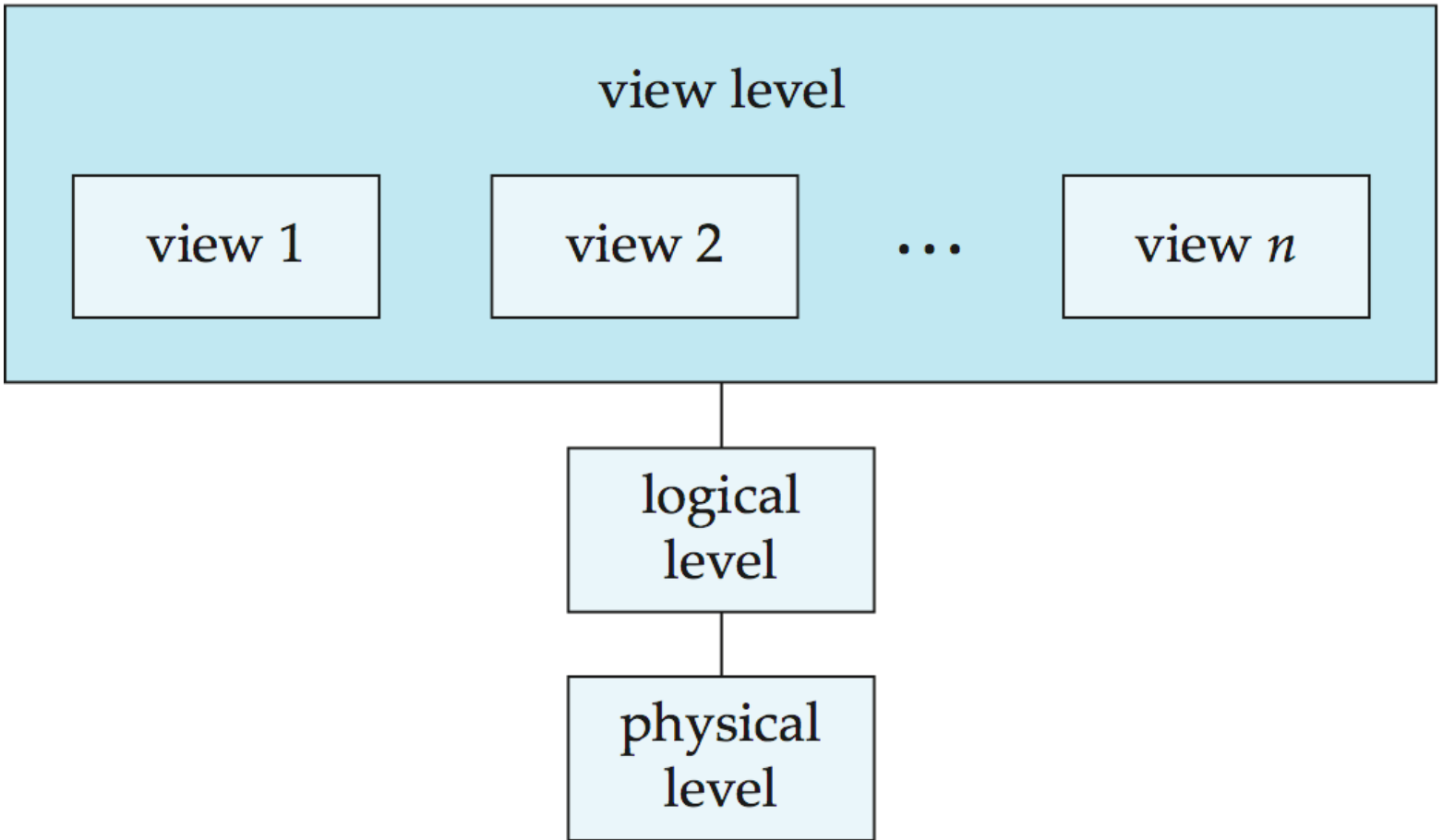
- **Nhắc lại**: Kiến trúc CDSL tập trung theo kiến trúc ANSI/SPARC



ES: External Schema
CS: Conceptual Schema
PS: Physical Schema

Kiến trúc D - DB

- Kiến trúc lược đồ ba mức ANSI/SPARC đáp ứng nhu cầu của một CSDL tập trung, *nhưng nó không đáp ứng đầy đủ cho một D-DB.*



Kiến trúc D - DB

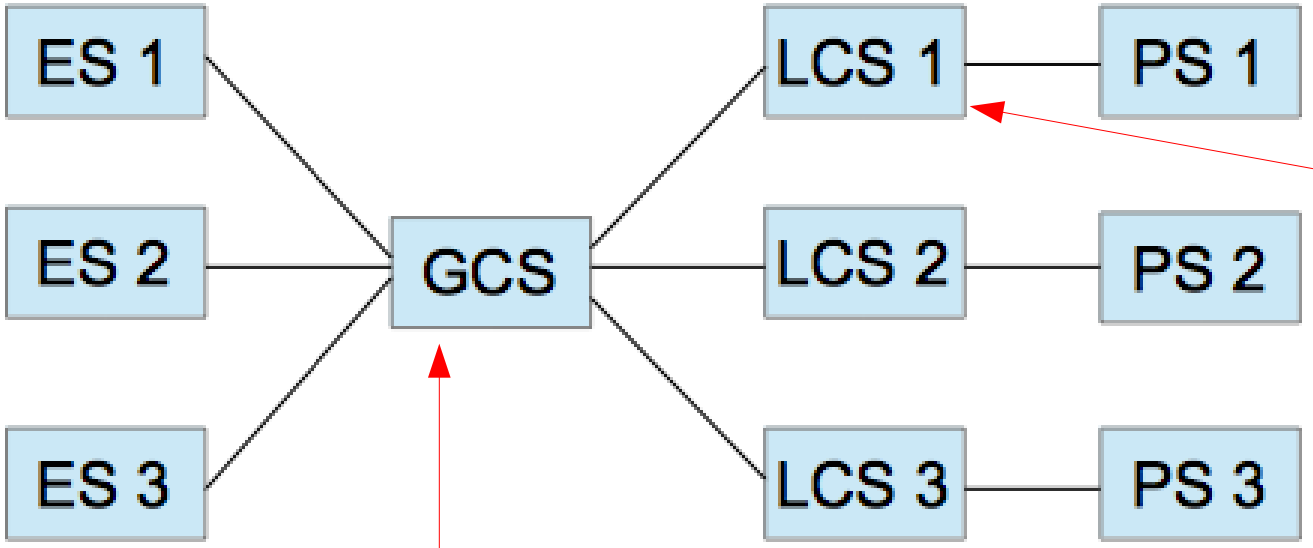
- Trong một D-DB, các khung nhìn của người dùng phải kết hợp thông tin trên các nút khác nhau.

  => khung nhìn của họ cần phải được xây dựng dựa trên một khung nhìn được tích hợp các lược đồ quan niệm cục bộ tham gia vào CSDL.

=> *Thêm một mức trừu tượng mới được gọi là lược đồ quan niệm tổng thể* (global conceptual schema GCS) vào kiến trúc lược đồ ba mức ANSI/SPARC.

Kiến trúc D - DB

Kiến trúc tham chiếu:



Các lược đồ quan niệm cục bộ cung cấp một khung nhìn cục bộ về dữ liệu được lưu trữ cục bộ tại mỗi nút.

- GCS là một khung nhìn tích hợp của tất cả các lược đồ quan niệm cục bộ (LCS).
- Nó cung cấp các thành phần cơ bản để tạo ra các khung nhìn bên ngoài (ES) cho người dùng của hệ thống phân tán.

- GCS: Global Conceptual Schema
- LCS: Local Conceptual Schema

Kiến trúc D - DB

- Trong một hệ thống phân tán mà mỗi nút là một DBMS sử dụng hệ thống quan hệ,
  - GCS cung cấp thông tin về các bảng, tất cả các khóa chính, tất cả các khoá ngoài, tất cả các ràng buộc, ...
  - GCS không chứa thông tin về nơi mà mỗi bảng được lưu trữ, làm thế nào mỗi bảng được phân đoạn, hoặc có bao nhiêu bản sao của mỗi đoạn có trong D-DB.

=> Vì vậy một số các thông tin (không có trong GCS) cần thêm vào để cung cấp sự trong suốt về vị trí, phân đoạn, và nhân bản.

- GCS với các thông tin tăng cường này được gọi là một **từ điển dữ liệu toàn cục (GDD - Global Data Dictionary).**

Từ điển dữ liệu toàn cục - GDD

- Là phần mở rộng của từ điển ANSI/SPARC, mô tả vị trí và tính chất của các thành phần của D-DB như các đoạn,...
- Là meta-database chứa các thông tin về database
- Thường gồm 5 phần...

Từ điển dữ liệu toàn cục - GDD

1. **GCS** chứa thông tin về

- Các bảng, các cột, các loại dữ liệu
- Các ràng buộc về khoá, ràng buộc cột ...

- GCS cung cấp sự độc lập giữa dữ liệu và ứng dụng, mà được yêu cầu bởi tất cả các hệ thống DBMS theo chuẩn ANSI/SPARC.

Từ điển dữ liệu toàn cục - GDD

2. Data Directory (Thư mục Dữ liệu - DD).

- DD chứa thông tin về vị trí của các đoạn dữ liệu.
  - Thông tin này cho biết vị trí của các site bằng cách chỉ ra URL, tên site, địa chỉ IP, v.v ...cho site chứa dữ liệu.
- DD cho phép D-DB cung cấp sự trong suốt về vị trí.

Từ điển dữ liệu toàn cục - GDD

3. Fragmentation Directory (Thư mục phân đoạn - FD).

- FD có thông tin về các đoạn dữ liệu trong hệ thống.
- FD thường chứa:
  - các điều kiện được sử dụng để tạo các phân đoạn ngang (horizontal fragment),
  - kết nối cột cho các phân đoạn dọc (vertical fragment),
  - các cột là thành phần của phân đoạn dọc,
  - khóa chính của các đoạn, ...
- Phần này của GDD cung cấp sự trong suốt về phân đoạn.

Từ điển dữ liệu toàn cục - GDD

4. Replication Directory (Thư mục nhân bản - RD).

- RD chứa thông tin về nhân bản.
  - Nó bao gồm số lượng bản sao cho mỗi bảng hoặc một đoạn của một bảng.
- Lưu ý: thông tin này kết hợp với thông tin DD là đủ để định vị tất cả các bản sao của bất kỳ một đoạn hoặc một bảng.
- RD cho phép D-DB cung cấp sự trong suốt trong nhân bản.

Từ điển dữ liệu toàn cục - GDD

5. Network Directory (Thư mục mạng - ND).

- ND có thông tin về:
  - hình trạng mạng (topology),
  - tốc độ truyền thông, ... cho tất cả các site tham gia vào D-DB.
- Phần này của GDD cho phép D-DB cung cấp sự trong suốt của mạng.

Sự trong suốt (transparency)

- Cài đặt hệ thống CSDL phân tán rất phức tạp
- Dấu các chi tiết cài đặt khỏi người dùng: người dùng không biết về sự phân tán dữ liệu => sự trong suốt.
- Các hình thức trong suốt:
  - Độc lập dữ liệu (data independence)
  - Trong suốt về mạng
  - Trong suốt trong nhân bản
  - Trong suốt trong phân đoạn
- Trong suốt có thể được cung cấp tại các mức khác nhau của hệ thống

Sự trong suốt (transparency)

- **Độc lập dữ liệu** (data independence)
  - Mức luận lý: ứng dụng của người dùng không bị ảnh hưởng khi có thay đổi về cấu trúc luận lý của CSDL
  - Mức vật lý: dấu đi các chi tiết về cấu trúc lưu trữ

Sự trong suốt (transparency)

- **Trong suốt về mạng** (network transparency)
  - Vị trí (location):
    - Mỗi thao tác trên dữ liệu là độc lập với cả hai vị trí và hệ thống nơi nó được thực thi
    - Người dùng truy xuất lược đồ quan niệm nhờ vào khung nhìn, họ không biết dữ liệu thực sự nằm trên site nào
  - Tên: tên duy nhất được cung cấp cho mỗi đối tượng CSDL

Sự trong suốt (transparency)

- **Trong suốt trong nhân bản**
  - Dữ liệu được nhân bản để tính đến độ tin cậy và hiệu suất
  - Người dùng phải không biết về sự tồn tại của các bản sao của dữ liệu

    => nếu có sự sửa đổi dữ liệu, hệ thống phải chịu trách nhiệm cập nhật tất cả các bản sao.

Sự trong suốt (transparency)

- **Trong suốt trong phân đoạn**
  - Các quan hệ của CSDL được chia thành các đoạn nhỏ hơn vì lý do hiệu suất và tính khả dụng
  - Các truy vấn toàn cục phải được dịch sang các truy vấn phân đoạn
  - Người sử dụng không nhận thức được sự tồn tại của các đoạn và làm việc trên các quan hệ toàn cục.

    => Vấn đề xử lý truy vấn.

Ôn tập ĐSQH