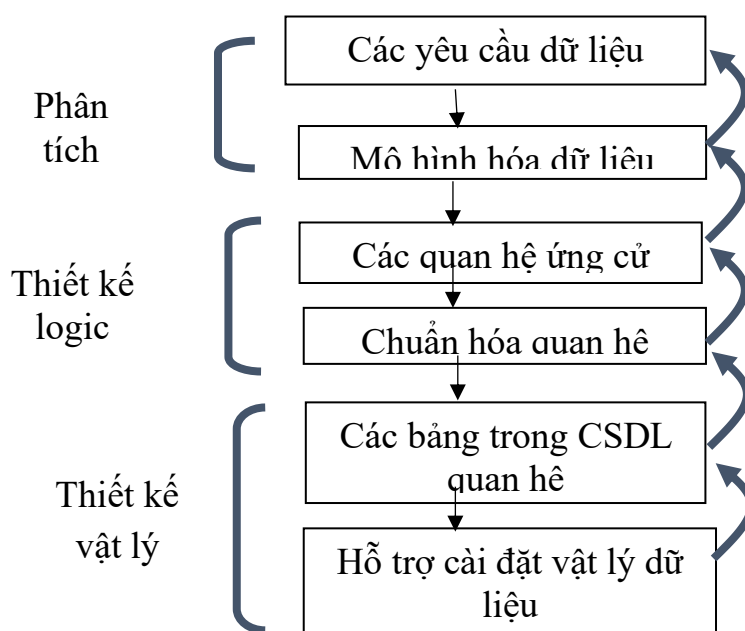


## BÀI 5. MÔ HÌNH HÓA DỮ LIỆU

- **Mục đích:** Cung cấp cho sinh viên kiến thức về thiết kế cơ sở dữ liệu mức vật lý.
- **Yêu cầu:** Bài học cung cấp cho SV các kiến thức về các bước tiến hành thiết kế CSDL, kỹ thuật chuyển từ mô hình CSDL logic sang mô hình CSDL vật lý
- **Hình thức tổ chức dạy học:** Lý thuyết, tự học
- **Thời gian:** Lý thuyết( trên lớp: 2; online: 2) Tự học, tự nghiên cứu:8
- **Nội dung chính:**

BÀI 5. MÔ HÌNH HÓA DỮ LIỆU .....	1
1. Các bước tiến hành thiết kế CSDL .....	1
2. Chuẩn hóa quan hệ.....	2
3. Các bảng trong CSDL quan hệ .....	2
4. Quy tắc chuyển từ mô hình CSDL logic sang mô hình CSDL vật lý.....	3
5. Xem xét tính hiệu quả.....	5
6. Thực thi bên trong và điều chỉnh .....	8

### 1. Các bước tiến hành thiết kế CSDL



- Phân tích dữ liệu yêu cầu:
  - Thực hiện đồng thời với mô hình hóa dữ liệu
  - Tìm hiểu về nghiệp vụ và các quy tắc nghiệp vụ

- Mô hình hóa dữ liệu:
  - Xây dựng mô hình thực thể liên kết
  - Xây dựng từ điển dữ liệu
- Xác định các quan hệ ứng cử: Chuyển từ mô hình thực thể liên kết thành mô hình quan hệ (bản ghi logic)
- Chuẩn hóa các quan hệ: Chuẩn hóa các quan hệ ít nhất về dạng chuẩn 3 (3NF) (bản ghi logic)
- Các bảng trong CSDL quan hệ: quyết định cấu trúc thực sự của bảng của CSDL quan hệ
- Hỗ trợ các cài đặt vật lý trong CSDL: Thiết kế các bảng dựa trên một hệ quản trị CSDL cụ thể

## 2. Chuẩn hóa quan hệ

- Sử dụng luật chuẩn hóa dữ liệu để kiểm tra các quan hệ ứng cử ở bước trên đã ở dạng chuẩn 3 (3NF) để đảm bảo không có dị thường dữ liệu (cách tiếp cận từ trên xuống)
- Ngoài ra chuẩn hóa được xem là cách tiếp cận từ dưới lên vì nó bắt đầu với một tập hợp các thuộc tính và tổ chức chúng thành các quan hệ có cấu trúc tốt hơn loại bỏ dư thừa dữ liệu (có thể loại bỏ dữ liệu trùng lặp mà không làm mất mát thông tin).
- Nguyên tắc chuẩn hóa:
  - Nhóm tất cả các thuộc tính liên quan vào trong một quan hệ
  - Áp dụng các luật chuẩn hoá để tách quan hệ đó thành các quan hệ có cấu trúc tốt hơn, giảm bớt dư thừa dữ liệu.

## 3. Các bảng trong CSDL quan hệ

Kết hợp kết quả từ hai cách tiếp cận theo nguyên tắc:

1. Kiểm tra sự thống nhất về tên gọi của các quan hệ và các thuộc tính trong hai cách tiếp cận. Nếu cùng tên nhưng khác nghĩa thì phải đặt lại tên cho khác nhau. Nếu cùng nghĩa nhưng khác tên thì phải đặt lại tên cho giống

---

nhau,

2. Lấy tất cả các quan hệ khác nhau từ hai kết quả. Với hai quan hệ trùng nhau thì lấy tất cả các thuộc tính có trong hai quan hệ từ hai kết quả. Kiểm tra lại quyết định thuộc tính nào phù hợp để lại trong mỗi thực thể. Loại bỏ các thuộc tính không cần thiết.

#### **4. Quy tắc chuyển từ mô hình CSDL logic sang mô hình CSDL vật lý**

CSDL được thiết kế ở mức logic là các quan hệ đã chuẩn hóa và độc lập với hệ quản trị cơ sở dữ liệu. Trong giai đoạn thiết kế vật lý cần chuyển từ mô hình dữ liệu sang mô hình cài đặt là các bảng quan hệ trong một HQTCSDDL thực sự (ví dụ như HQTCSDDL MS SQL Server).

- a. Mỗi quan hệ biểu diễn thành 1 bảng tương ứng. Với mỗi bảng cần chỉ rõ
  - Khóa chính
  - Tất cả các thuộc tính mô tả
- b. Mỗi thuộc tính sẽ được biểu diễn bởi một trường trong bảng. Mỗi trường cần có:
  - Một tên duy nhất (trong bảng chứa nó). Tên trường cần ngắn gọn nhưng đủ nghĩa
  - Một mô tả (trong từ điển dữ liệu)
  - Một kiểu dữ liệu và kích thước (VD: integer, char, date, logical,...).Chọn kiểu DL thỏa mãn:
  - Tối thiểu không gian lưu trữ
  - Trình bày tất cả các giá trị của trường
  - Tăng tính toàn vẹn cho dữ liệu
  - Hỗ trợ tất cả các thao tác của DL
- Trường tùy chọn hay bắt buộc. Nếu một trường là bắt buộc, người dùng cần phải cung cấp giá trị cho nó trong mỗi dòng khi CSDL được cập nhật => Nó cần phải thiết lập NOT NULL. Nếu một trường là tùy chọn, nó có thể chứa giá trị NULL

- Giá trị hợp lệ và giá trị mặc định: HQTCSDL cho phép thiết lập giá trị hợp lệ và giá trị mặc định cho từng trường. Ví dụ: Số điện thoại được định nghĩa là một dãy gồm 12 ký tự chữ số (tính hợp lệ). Giá trị mặc định cho ngày lập hóa đơn là ngày hiện thời
  - Thuận lợi của việc chỉ rõ tiêu chuẩn xác nhận tính hợp lệ là chúng được thi hành tự động và không cần phải gắn vào tất cả các chương trình ứng dụng.
  - Giá trị mặc định có thể được sử dụng trong trường hợp người dùng không nhập dữ liệu vào. Nó có thể làm cho việc nhập dữ liệu nhanh hơn và giảm lỗi

c. Khóa chính và khóa ngoài

- Hầu hết các HQTCSDL hiện đại cho phép chỉ rõ khóa chính và khóa ngoài khi thiết lập định nghĩa bảng. Nếu HQTCSDL không thiết lập tự động thì phải định nghĩa khóa chính có tính duy nhất và NOT NULL
- Quan hệ giữa các bảng được thiết lập bởi khóa ngoài.
- Cần phải đảm bảo rằng buộc toàn vẹn tham chiếu được thi hành.

d. Khóa giả

- Các khóa tốt nhất là nên ngắn để cho việc tìm kiếm nhanh hơn, tốn ít không gian và dữ liệu nhập vào ít lỗi hơn.
- Nếu một số khóa là tương đối dài (thường là 3 thuộc tính trở lên) cần xem xét thay thế chúng bằng một khóa giả.
- Có hai vấn đề khi sử dụng khóa giả là:
  - Cần thiết lập một kỹ thuật sinh ra giá trị duy nhất cho khóa giả (sinh tự động)
  - Cần cung cấp một **ràng buộc duy nhất** cho một định danh ‘tự nhiên’, để tránh trường hợp một vài bản ghi của một thực thể lưu dữ liệu giống nhau nhưng giá trị khóa giả lại khác nhau

---

## 5. Xem xét tính hiệu quả

### 5.1. Tiêu chí đánh giá sự thi hành CSDL:

- Tốc độ truy cập dữ liệu
- Khả năng “phục vụ” đồng thời
- Tính sẵn sàng của hệ thống
- Dung lượng lưu trữ dữ liệu
- Sử dụng tài nguyên hệ thống
- Bộ nhớ
- CPU
- Băng thông

### 5.2. Sử dụng chỉ số (Index)

Trong giai đoạn thiết kế, kỹ thuật thiết lập các chỉ số là kỹ thuật quan trọng nhất giúp cho định vị và truy cập dữ liệu vào các bảng CSDL nhanh nhất.

- Một nguyên tắc chung là, nên có :
  - Một chỉ số cho mỗi khóa chính
  - Một chỉ số cho mỗi khóa ngoài
  - Các trường thường xuyên được truy nhập cần phải có chỉ số
  - Với chỉ số trên các trường ghép cần phải cung cấp thành phần dẫn dắt nếu không chỉ số này sẽ không được sử dụng
- Lợi ích của các chỉ số
  - Tìm kiếm và khôi phục các bản ghi nhanh hơn
  - Hiệu quả trong việc lấy dữ liệu từ nhiều bảng
  - Có thể có các thứ tự sắp xếp khác nhau trên một bảng
  - Ngăn cản các đầu vào trùng lặp nếu chỉ số là duy nhất
  - Nhiều mức sắp xếp được thiết lập trên một chỉ số của các trường ghép
- Nhược điểm của việc dùng chỉ số
  - Cần phải có thêm không gian nhớ trên bộ nhớ thứ cấp (đĩa) để lưu

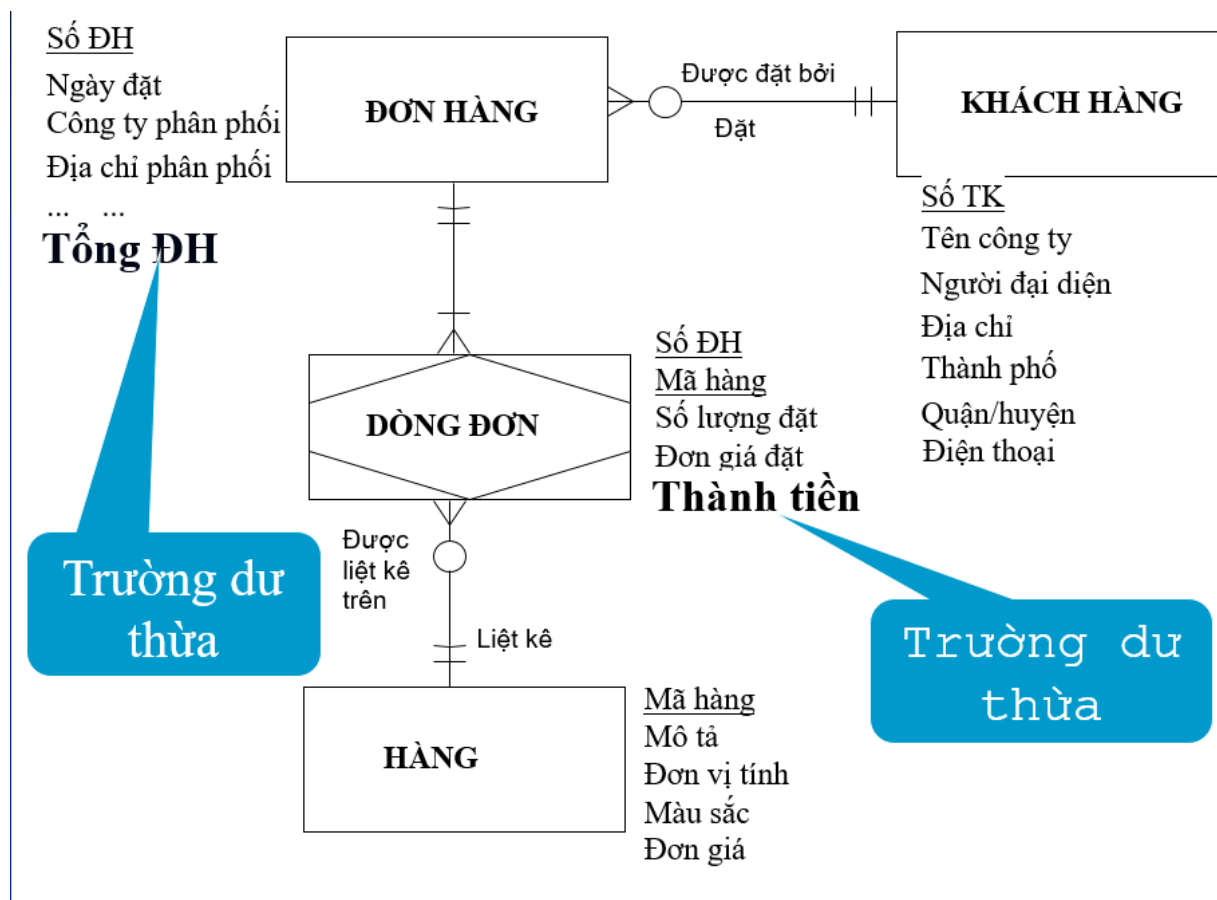
trữ chỉ số

- Tăng thời gian để cập nhật tất cả các chỉ số khi các bản ghi dữ liệu được cập nhật.

### 5.3. Phá vỡ chuẩn

- Phá vỡ chuẩn có nghĩa là cố ý đưa các dư thừa vào trong CSDL để cải tiến tốc độ thực hiện, và điều này có nghĩa là phá vỡ các quan hệ đã chuẩn hóa.
- Khi xem xét tính hiệu quả của việc truy nhập dữ liệu, trong quá trình xây dựng các bảng từ các quan hệ đã chuẩn hóa
  - Chúng ta có thể quyết định ghép bảng hoặc tách bảng
  - Chúng ta có thể quyết định lặp lại các trường hay đưa vào các trường suy diễn (VD: thành tiền, tổng)

### Ví dụ 1: Đưa thêm thuộc tính suy diễn

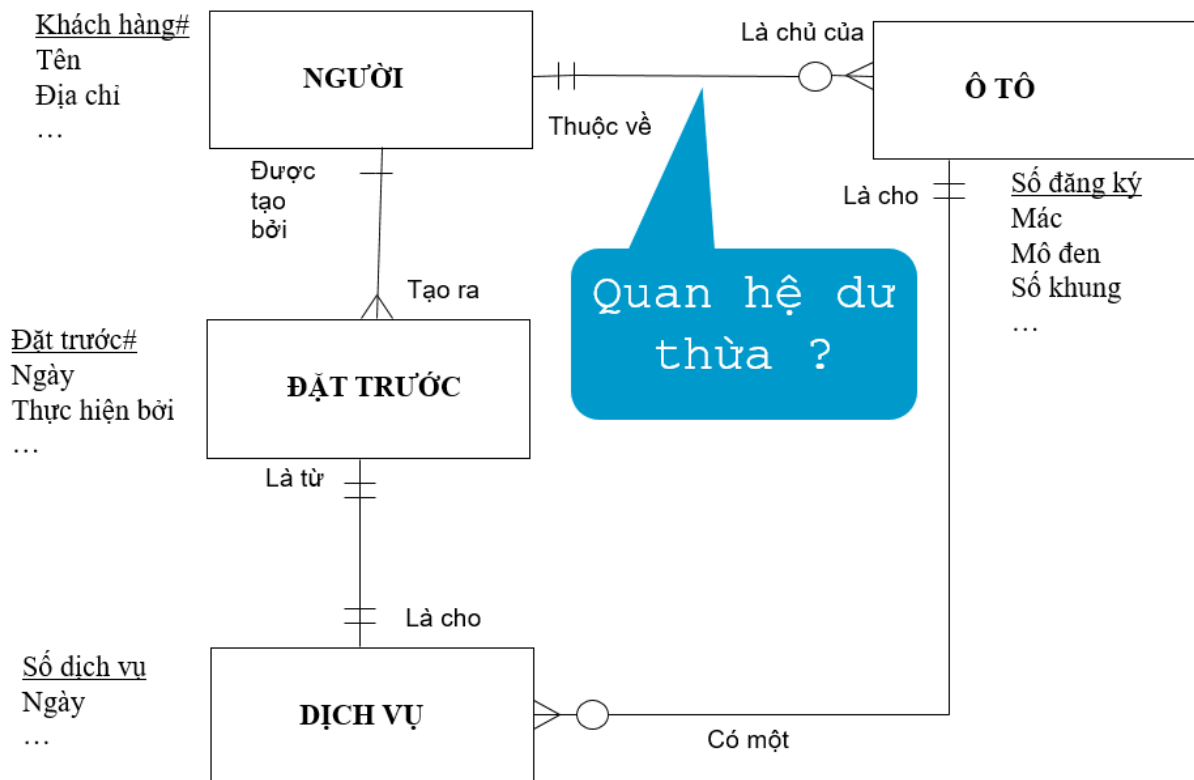


- Các quan hệ đã chuẩn hóa không có các thuộc tính tính Thành tiền. Tuy nhiên tất cả những lần truy nhập đơn hàng thì tổng các dòng hàng đều

được tính lại. Bằng cách lưu Tổng ĐH và Thành tiền tương ứng trong ĐƠN HÀNG và DÒNG HÀNG, sẽ giảm bớt sự tính toán và tăng tốc độ truy nhập.

- Tuy nhiên có một bất lợi trong việc bảo trì. Bạn cần đảm bảo rằng tất cả những lần bạn thay đổi Số lượng đặt hay đơn giá bạn đều cần phải tính lại tổng tiền.

## **Ví dụ 2: Quan hệ vòng**



- Từ quan điểm thực thi, bất cứ một truy vấn CSDL nào yêu cầu truy nhập hai bảng không có quan hệ trực tiếp đều có thể thực hiện được theo con đường gián tiếp thông qua một bảng khác.
- Nếu quan hệ NGƯỜI-ÔTÔ bị gỡ bỏ, công ty sửa chữa ô tô vẫn xác định được người nào làm chủ xe nào nhưng yêu cầu này sẽ cần đến nhiều truy nhập CSDL hơn

### **5.4. Các tính hiệu quả khác**

- Nếu bạn cần lưu trữ nhiều quy tắc và các mô tả quy tắc khác nhau, hãy đặt

chúng vào trong một bảng tham chiếu chung. Ứng dụng sẽ đọc bảng này khi khởi động và sử dụng dữ liệu trong suốt thời gian chạy mà không cần truy nhập bảng nữa.

- Khi thực hiện kết nối bảng trong truy vấn, việc chỉ rõ thứ tự kết nối của các bảng thường là quan trọng. VD giả sử bạn cần một tập con dữ liệu từ hai bảng khác nhau và cần kết nối hai bảng đó lại. Sẽ hiệu quả hơn nếu lọc ra các tập con dữ liệu cần thiết từ mỗi bảng, sau đó thực hiện kết nối các tập con đó (thay vì kết nối cả hai bảng lớn), sau đó lọc ra tập con dữ liệu mà bạn thực sự cần.
- Nếu một truy vấn cụ thể được thực thi quá chậm, hãy sử dụng “vết”, nếu có sẵn, tìm xem có bao nhiêu truy vấn đã được thực hiện và sử dụng thông tin này để sửa chữa trình tự các lệnh cải tiến tốc độ
- Tách các bảng lớn khi một số trường được truy nhập rất thường xuyên còn một số trường khác thỉnh thoảng mới được truy nhập.
- Kết quả là khi bạn thiết kế logic có thể chỉ có một thực thể (và do đó có một quan hệ) với nhiều thuộc tính, nhưng CSDL vật lý có thể được cài đặt với hai hay nhiều hơn các bảng liên quan với nhau trên cơ sở quan hệ một-một. Trong đó một bảng có các trường thường xuyên được truy nhập, trong khi các bảng khác sẽ có các trường ít khi được truy nhập.

## **6. Thực thi bên trong và điều chỉnh**

- Khi cài đặt cơ sở dữ liệu trong một hệ QTCSDL lớn, việc thực thi và điều chỉnh bên trong là nhiệm vụ của người quản trị CSDL (DBA).
- Các công việc của thực thi và điều chỉnh bên trong bao gồm:
  - Tổ chức các thủ tục sao lưu và phục hồi cho CSDL
  - Tổ chức và bảo trì sự an toàn bảo mật cho CSDL
  - Thiết lập các ràng buộc toàn vẹn trong CSDL
  - Xác định không gian CSDL cho các bảng và chỉ số



---

**Yêu cầu sinh viên chuẩn bị:**

Đọc trước đề cương bài giảng chi tiết và slides bài giảng, xem video bài giảng, làm bài trắc nghiệm bài 7.