Các mô hình dữ liệu

Nguyễn Thị Oanh
oanhnt@soict.hust.edu.vn

Bộ môn Hệ thống thông tin (http://is.hust.edu.vn/)
Viện CNTT&TT
Trường Đại học Bách Khoa Hà Nội

Nội dung

- Tổng quan các mô hình dữ liệu
- Mô hình dữ liệu quan hệ

Mô hình dữ liệu

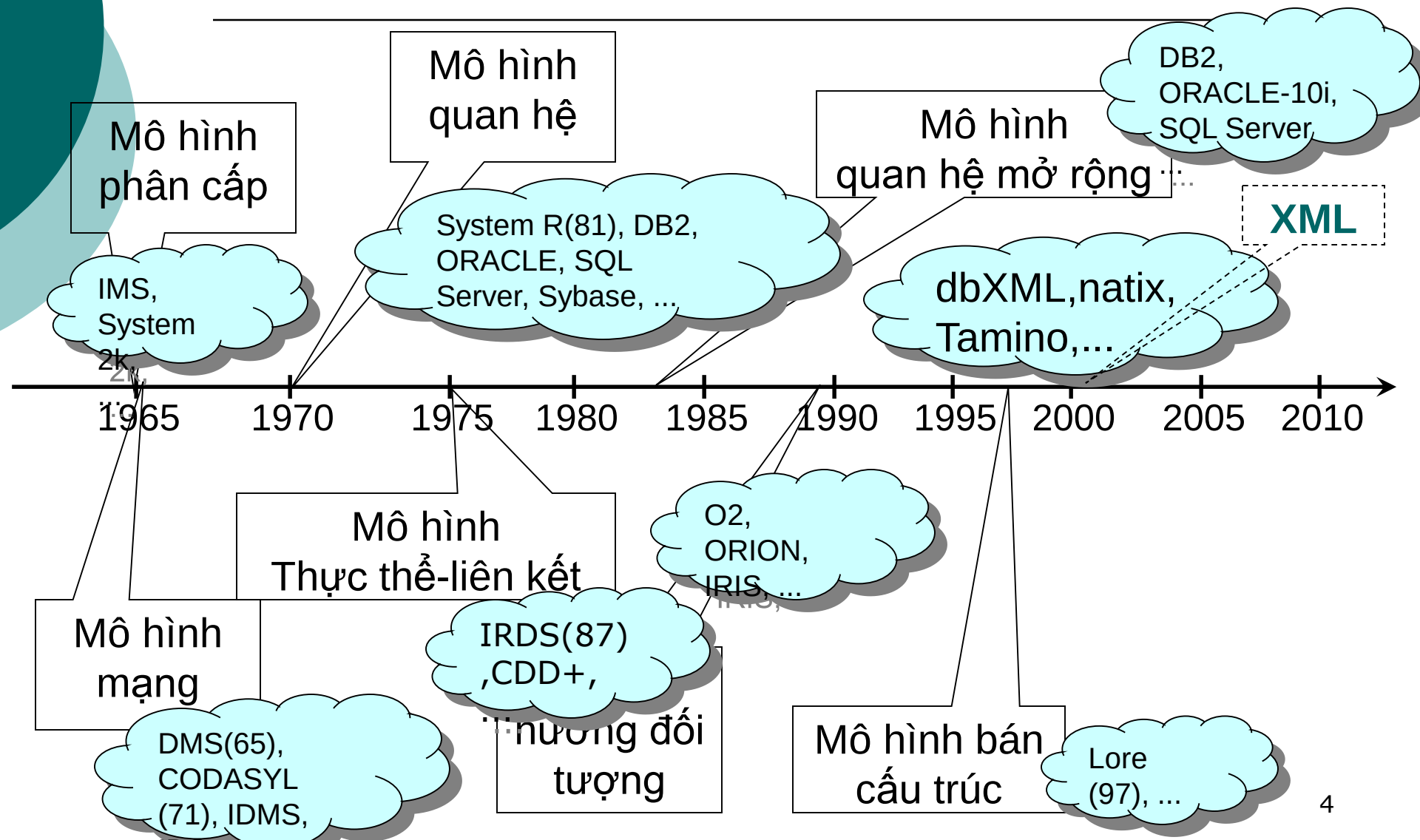
*"A data model is a plan for building a database"**

- Mô hình dữ liệu gồm [Codd, 1980]
  - Một tập hợp các cấu trúc của dữ liệu
  - Một tập các phép toán để thao tác với các dữ liệu
  - Một tập các ràng buộc về dữ liệu

Ví dụ: mô hình mạng, mô hình phân cấp, mô hình quan hệ, mô hình thực thể-liên kết, mô hình hướng đối tượng

*http://www.computerworld.com/databasetopics/data/story/0,10801,80205,00.html

Vài nét về lịch sử



Một vài mô hình dữ liệu

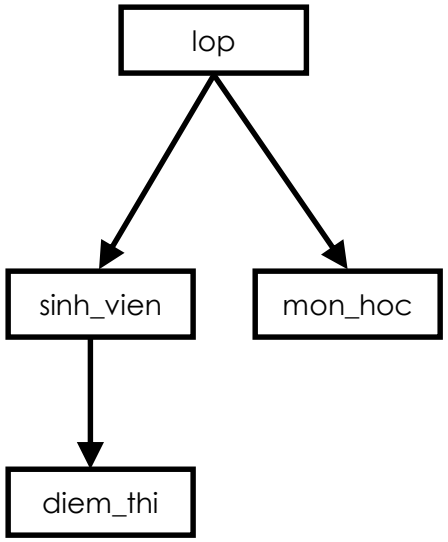
- Mô hình phân cấp
- Mô hình mạng
- Mô hình quan hệ
- Mô hình thực thể - liên kết
- Mô hình hướng đối tượng
- *Mô hình bán cấu trúc*
- *Mô hình dữ liệu của XML*

Đặt vấn đề

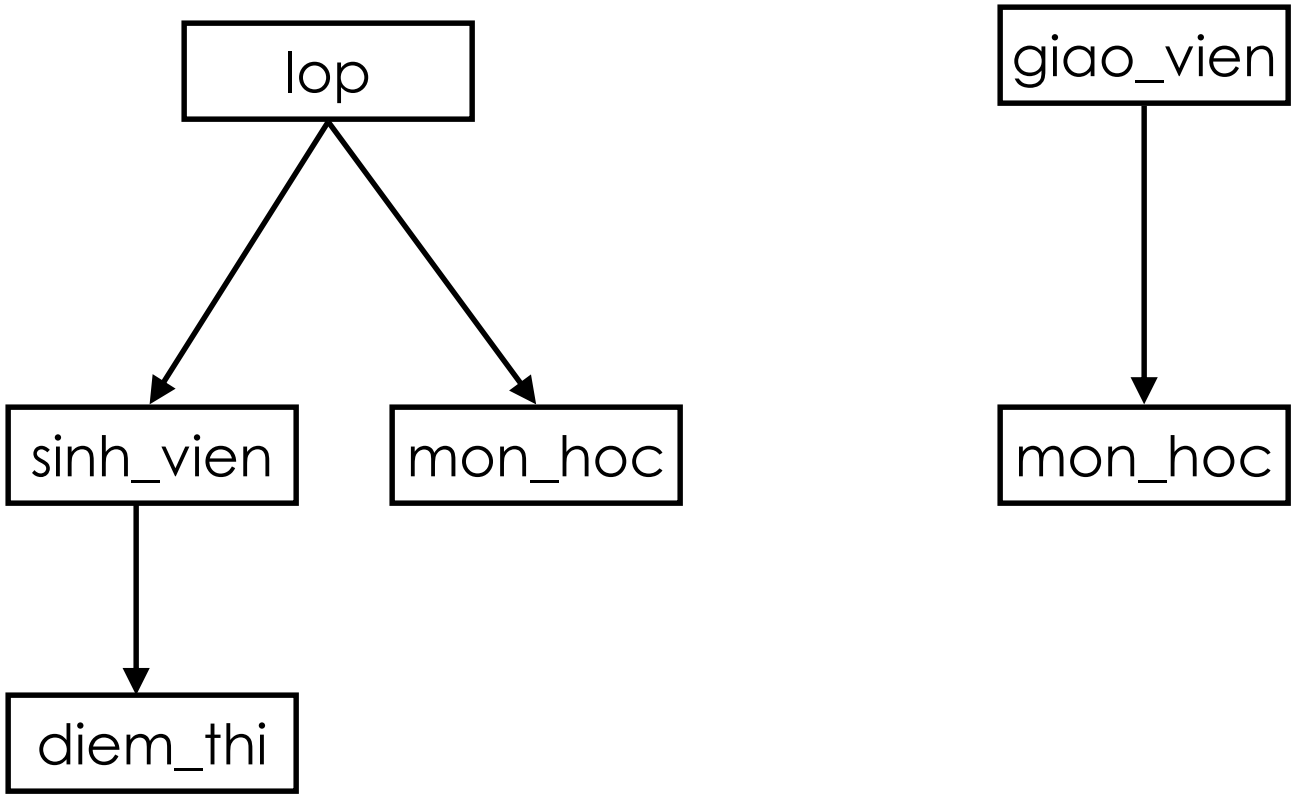
- Đặc điểm của các mô hình dữ liệu?
- Sự khác nhau giữa các mô hình dữ liệu?
- Các mô hình dữ liệu phổ biến ngày nay

Mô hình dữ liệu phân cấp (*Hierarchical data model*)

- Sự ra đời
  - Khoảng năm 60-65
- Biểu diễn: bằng cây
  - Quan hệ cha/con
  - Mỗi nút có một cha duy nhất
  - 1 CSDL = tập các cây
- Các khái niệm cơ bản
  - Bản ghi
  - Móc nối
    - Các phép toán: GET, GET UNIQUE, GET NEXT, GET NEXT WITHIN PARENT, ...



Ví dụ

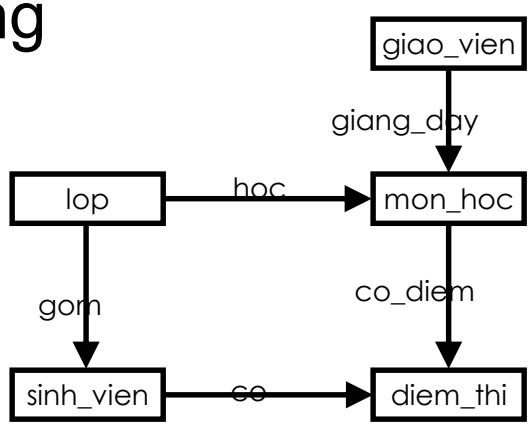


Nhận xét

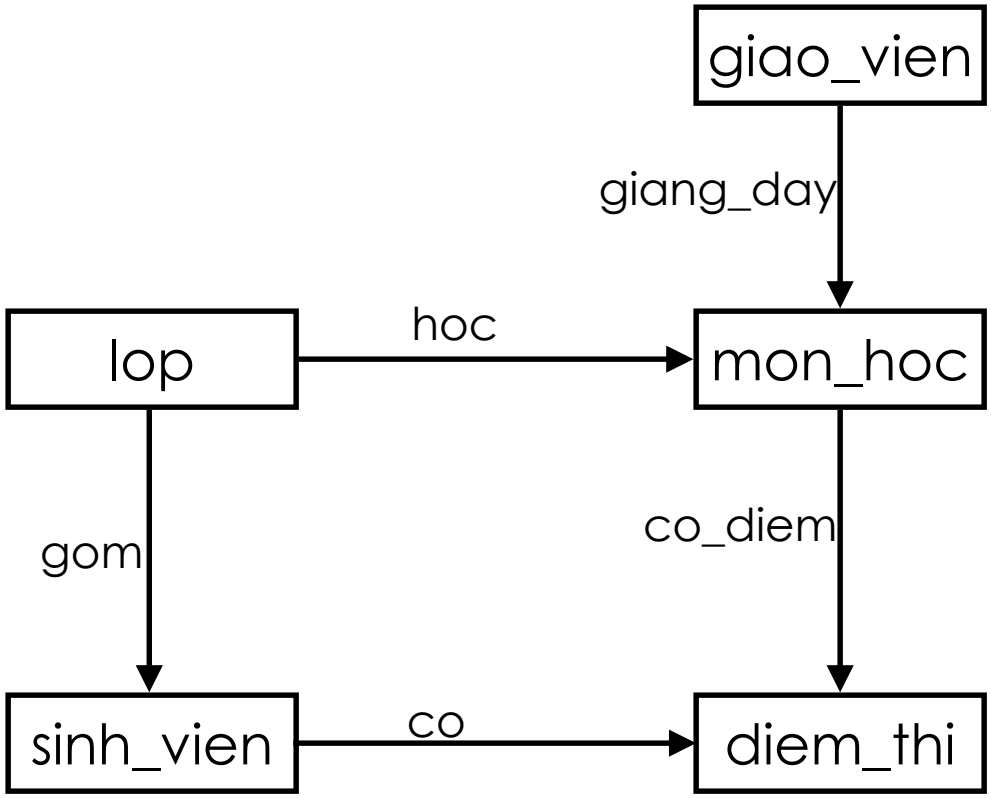
- Ưu điểm
  - Dễ xây dựng và thao tác
  - Tương thích với các lĩnh vực tổ chức phân cấp (vd: tổ chức nhân sự trong các đơn vị, ...)
  - Ngôn ngữ thao tác đơn giản (duyệt cây)
- Nhược điểm
  - Sự lặp lại của các kiểu bản ghi $\rightarrow$ dư thừa dữ liệu và dữ liệu không nhất quán
    - Giải pháp: bản ghi ảo
  - Hạn chế trong biểu diễn ngữ nghĩa của các móc nối giữa các bản ghi (chỉ cho phép quan hệ 1-n)

Mô hình dữ liệu mạng (*Network data model*)

- Sự ra đời
  - sử dụng phổ biến từ những năm 60, được định nghĩa lại vào năm 1971
- Biểu diễn: bằng đồ thị có hướng
- Các khái niệm cơ bản
  - Tập bản ghi (*record*)
    - Kiểu bản ghi (*record type*)
    - Các trường (*field*)
  - Móc nối (*link*)
    - Tên của móc nối
    - chủ (*owner*) – thành viên (*member*): theo hướng của móc nối
    - Kiểu móc nối: 1-1, 1-n, đệ quy
  - Các phép toán
    - Duyệt: FIND, FIND member, FIND owner, FIND NEXT
    - Thủ tục: GET



Ví dụ


giao_vien
giang_day
lop
hoc
mon_hoc
gom
co_diem
sinh_vien
co
diem_thi

Nhận xét

- Ưu điểm
  - Đơn giản
  - Có thể biểu diễn các ngữ nghĩa đa dạng với kiểu bản ghi và kiểu móc nối
  - Truy vấn thông qua phép duyệt đồ thị (*navigation*)
- Nhược điểm
  - Số lượng các con trỏ lớn
  - Hạn chế trong biểu diễn ngữ nghĩa của các móc nối giữa các bản ghi

Mô hình dữ liệu quan hệ (*Relational data model*)

- Sự ra đời
  - vào năm 1970[Codd, 1970]
- Biểu diễn: dưới dạng bảng
- Các khái niệm cơ bản
  - Thuộc tính: một tính chất riêng biệt của một đối tượng
    - Tên
    - Kiểu, miền giá trị
  - Quan hệ: được định nghĩa trên một tập các thuộc tính
  - Bộ giá trị: các thông tin của một đối tượng thuộc quan hệ
  - Khoá:
  - Các phép toán: hợp, giao, tích đề-các, lựa chọn, chiếu, kết nối, ...

Ví dụ

MON_HOC

maMH	tenmon	soHT
CNTT01	Nhập môn CSDL	4
CNTT02	Truyền DL và mạng	4
CNTT03	Phân tích và thiết kế hệ thống	4
HTTT01	Quản lý dự án	3

LOP

malop	lop	khoa	GVCN	loptruong
IT4	Tin 4	CNTT	Ng. V. Anh	Trần T. Bình
IT5	Tin 5	CNTT	Lê A. Văn	Ng. Đ. Trung
IT6	Tin 6	CNTT	Ng. T. Thảo	Trần M. Quế
IT7	Tin 7	CNTT	Ng. V. Quý	Ng. T. Phương

SINH_VIEN

maSV	tenSV	ngaysinh	nam	diachi	lop
SV0011	Trần T. Bình	1/4/1981	0	21 T. Q. B	IT4
SV0025	Ng. Đ. Trung	3/2/1980	1	56 Đ. C. V	IT5
SV0067	Trần M. Quế	26/3/1982	0	45 H. B. T	IT6
SV0034	Ng. T. Phương	29/2/1980	0	86 L. T. N	IT7

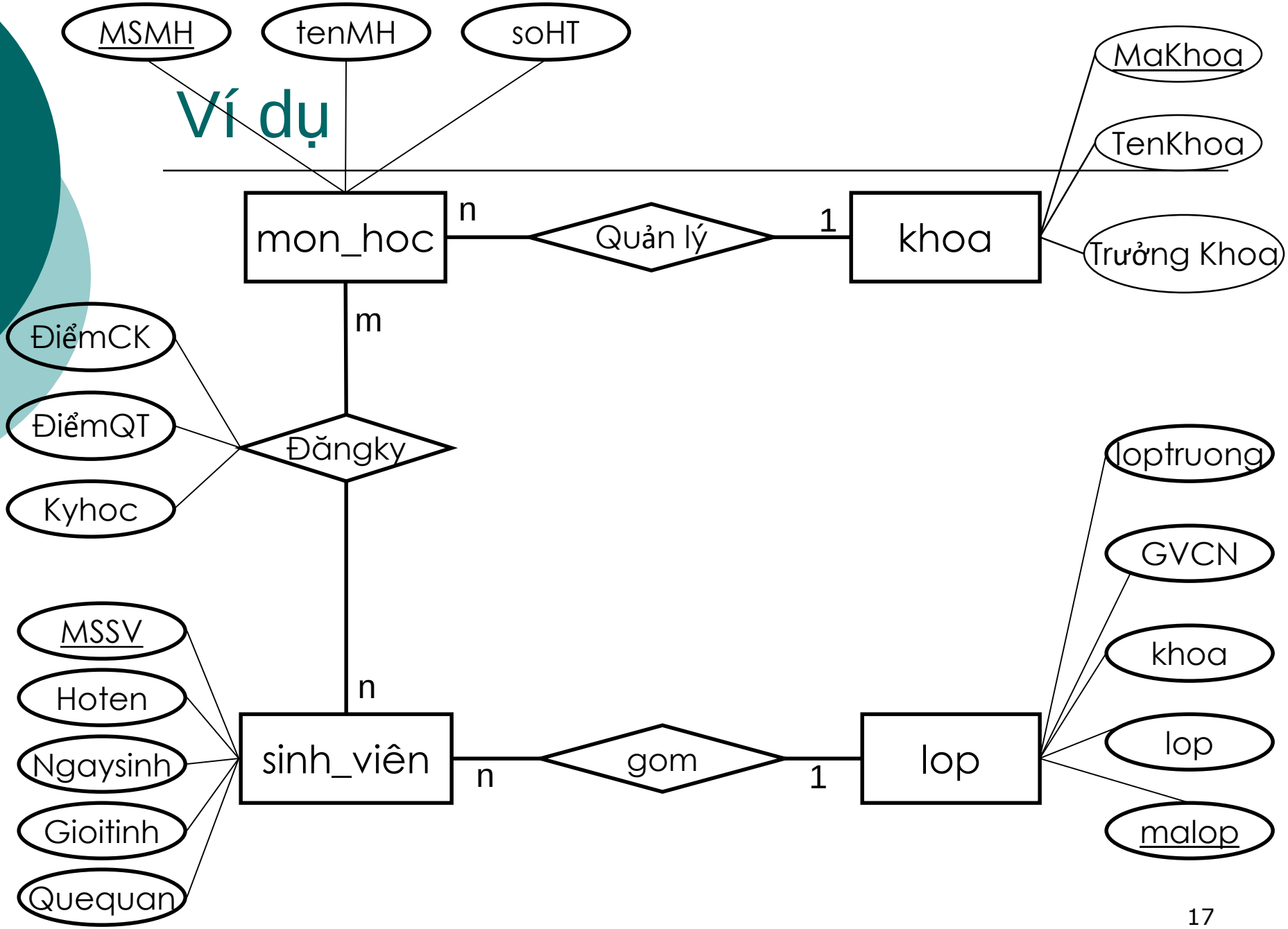
Nhận xét

- Ưu điểm
  - Dựa trên lý thuyết tập hợp
  - Khả năng tối ưu hoá các xử lý phong phú
- Nhược điểm
  - Hạn chế trong biểu diễn ngữ nghĩa
  - Cấu trúc dữ liệu không linh hoạt

Mô hình dữ liệu thực thể - liên kết (*Entity-Relational data model*)

- Sự ra đời
  - Xuất phát từ nhu cầu mô hình hoá ngữ nghĩa dữ liệu và phát triển phần mềm
  - đề xuất 1975 [Chen, 1976] [Chen, 2002]
- Biểu diễn: bằng sơ đồ thực thể - liên kết
- Các khái niệm cơ bản
  - Thực thể: một đối tượng trong thế giới thực
  - Thuộc tính: một đặc tính của một tập thực thể
  - Khoá:xác định sự duy nhất của 1 thực thể
  - Liên kết: mối liên hệ có nghĩa giữa nhiều thực thể
    - Mỗi liên kết có thể có các thuộc tính
    - 1-1, 1-n, n-m, đệ quy

Ví dụ


MSMH
tenMH
soHT
MaKhoa
TenKhoa
Trưởng Khoa
mon_hoc
n
Quản lý
1
khoa
m
ĐiểmCK
ĐiểmQT
Kyhoc
Đăngky
loptruong
GVCN
khoa
lop
malop
MSSV
Hoten
Ngaysinh
Gioitinh
Quequan
n
sinh_viên
n
gom
1
lop

Nhận xét

- Ưu điểm
  - dễ dàng biểu diễn cái mà con người nhận thức từ thế giới thực
  - Biểu diễn ngữ nghĩa phong phú của các thực thể và quan hệ giữa các thực thể
- Nhược điểm
  - Không dễ dàng ánh xạ vào những cấu trúc lưu trữ trên máy tính

Mô hình dữ liệu hướng đối tượng (*Object-oriented data model)*

- Sự ra đời
  - Khoảng đầu những năm 90
- Biểu diễn: sơ đồ lớp
- Các khái niệm cơ bản
  - Đối tượng: một đối tượng trong thế giới thực, được xác định bởi một định danh duy nhất
  - Thuộc tính: biểu diễn một đặc tính của đối tượng,
  - Phương thức : thao tác được thực hiện trên đối tượng.
    - Tất cả các truy nhập vào thuộc tính của đối tượng đều phải được thực hiện thông qua các phương thức này.
  - Lớp: một cách thức để khai báo một tập các đối tượng có chung một tập thuộc tính và phương thức

Ví dụ

```
class sinh_vien {
        string maSV;
        string tenSV;
        date ngaysinh;
        boolean nam;
        string diachi;
        string lop;


        string ten();
        string ngay_sinh();
        string dia_chi();
        string lop();
        void gan_DC(string DC_moi);
        void gan_lop(string lop);
}
```

Ví dụ

```
class lop {
        string tenlop;
        string khoa;
}

class sinh_vien {
        string maSV;
        string tenSV;
        date ngaysinh;
        boolean nam;
        string diachi;
        lop lop_hoc;
}
```

Nhận xét

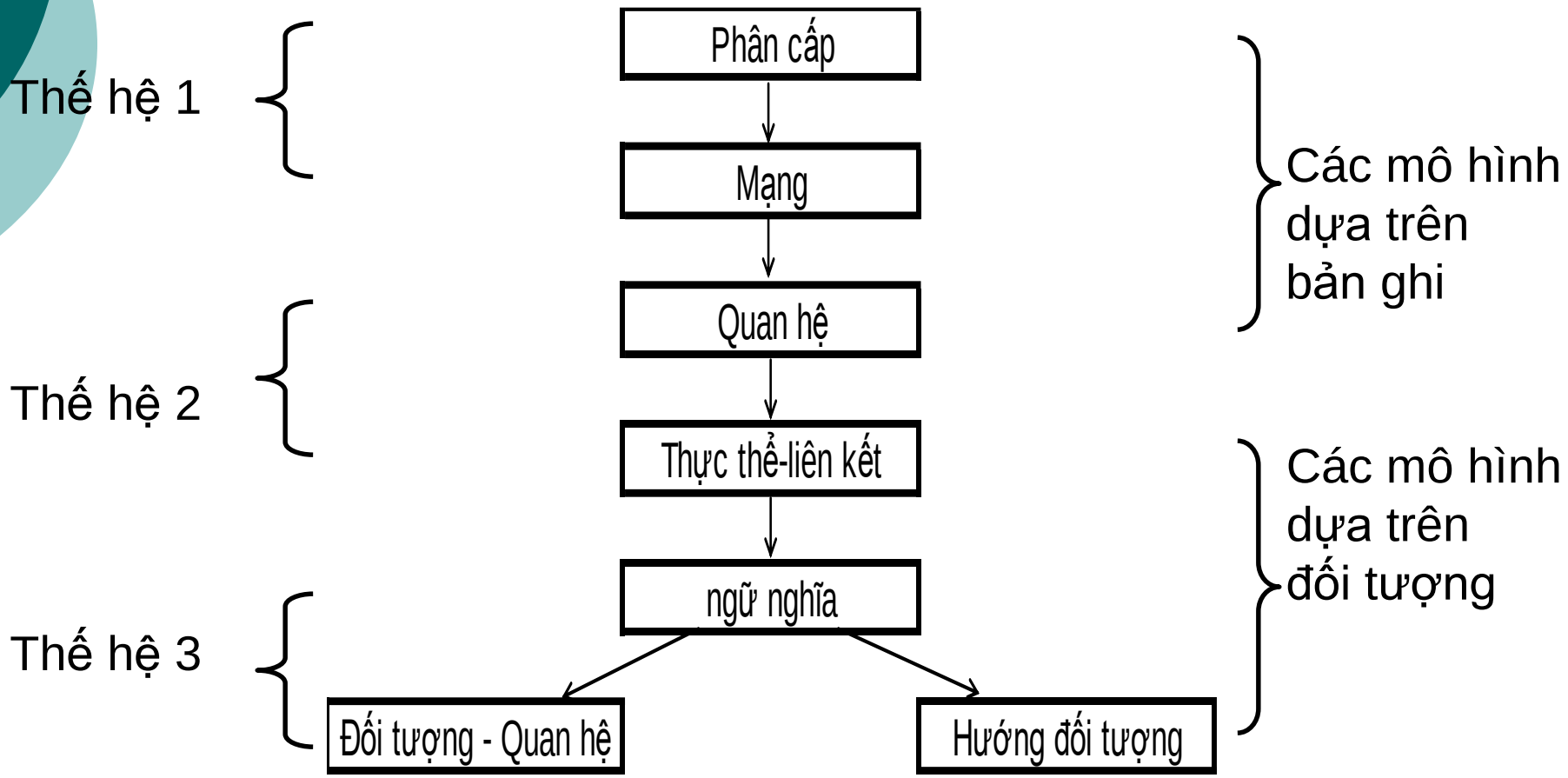
- Ưu điểm
  - Cho phép định nghĩa kiểu đối tượng phức tạp
  - Tính chất: bao đóng (*encapsulation*), kế thừa (*heritage*), đa hình (*polymorphism*)
- Nhược điểm
  - Cấu trúc lưu trữ phức tạp và có thể sử dụng nhiều con trỏ
  - Khả năng tối ưu hoá các xử lý bị hạn chế trong nhiều trường hợp

So sánh và đánh giá

Nhắc lại: Mô hình dữ liệu là một tập hợp các khái niệm dùng để mô tả cấu trúc của một CSDL

	Mô hình mạng	**Mô hình phân cấp**	**Mô hình quan hệ**	**Mô hình TT-LK**	**Mô hình HĐT**
biểu diễn ngữ nghĩa DL	hạn chế	hạn chế	***tương đối đa dạng***	đa dạng	đa dạng
lưu trữ DL	s/d nhiều con trỏ	dữ liệu lặp lại	dễ dàng và hiệu quả	khó lưu trữ	***cấu trúc phức tạp***
khả năng truy vấn	đơn giản	đơn giản	đa dạng		đa dạng
hiệu quả của truy vấn	ít khả năng tối ưu	ít khả năng tối ưu	tối ưu hoá tốt	không được xem xét (không hiệu quả)	***không h/q khi s/d nhiều con trỏ***

Phân loại các mô hình



Biến đổi giữa các mô hình dữ liệu

- Yêu cầu
  - Chuyển một sơ đồ dữ liệu từ một mô hình dữ liệu sang một mô hình khác
  - Đảm bảo tính « tương đương » của sơ đồ dữ liệu nguồn và đích
- Các biến đổi tương đương giữa các mô hình
  - Thực thể/liên kết - mạng
  - Thực thể/liên kết - hướng đối tượng
  - Quan hệ - hướng đối tượng

Mô hình dữ liệu quan hệ

Đặc điểm

- Dựa trên lý thuyết tập hợp
- dễ dàng ánh xạ đến cấu trúc lưu trữ vật lý
- Các khái niệm cơ bản
  - Thuật ngữ toán học: quan hệ, bộ và thuộc tính
  - Thuật ngữ hướng dữ liệu: bảng, bản ghi và trường
- Được biểu diễn bởi sơ đồ (lược đồ) quan hệ

Thuộc tính - trường

- Đ/n: là một tính chất riêng biệt của một đối tượng cần được lưu trữ trong CSDL để phục vụ cho việc khai thác dữ liệu về đối tượng
  - Ký hiệu: A
- Tên thuộc tính:
  - maSV, tenSV,ngaysinh,nam,diachi,lop
- Kiểu dữ liệu, miền giá trị (Dom(A))
  - text, number, boolean, date/time, memo
    - maSV: text(10)
    - nam: boolean
  - Dom(A): tập tất cả các giá trị A có thể nhận
    - Dom(maSV) = {text(10)}
    - Dom(nam) = {0 ,1}

Quan hệ - bảng

- Đ/n: được xác định trên một tập các thuộc tính A_i
- Ký hiệu:

$R(A_1,A_2, \dots A_n)$

SINH_VIEN (maSV,tenSV,ngaysinh, nam, diachi,lop)
LOP(malop,ten,khoa)

$R(A_1,A_2, \dots A_n) \subseteq Dom(A_1) \times \dots \times Dom(A_n)$

- **Tân từ**: quy tắc để xác định mối quan hệ giữa các thuộc tính A_i

$\forall lop \in SINH\_VIEN[lop], \ \exists malop \in LOP[malop] : lop = malop$

Bộ - bản ghi

- Đ/n: các thông tin của một đối tượng thuộc quan hệ
- Ký hiệu

  $t(a_1,a_2, \ldots a_n)$

  $t(a_1,a_2, \ldots a_n) \in Dom(A_1) \times \ldots \times Dom(A_n)$

SV0011	Trần T. Bình	1/4/1981	0	21 T. Q. B	IT4
SV0025	Ng. Đ. Trung	3/2/1980	1	56 Đ. C. V	IT5
SV0067	Trần M. Quế	26/3/1982	0	45 H. B. T	IT6
SV0034	Ng. T. Phương	29/2/1980	0	86 L. T. N	IT7

Sơ đồ quan hệ

- **Sơ đồ quan hệ (S)**: là sự trừu tượng hoá của quan hệ ở mức độ cấu trúc của một bảng 2 chiều
  - $S = \{A_i\}$
  - SINH_VIEN(maSV,tenSV, ngaysinh, nam, diachi,malop)
    LOP(malop,lop, khoa, GVCN, loptruong)
    MON_HOC(maMH,tenmon,soHT)
    ...
- **Thể hiện của quan hệ**: tập hợp các bộ giá trị của quan hệ R vào một thời điểm
- **Sơ đồ cơ sở dữ liệu quan hệ:** tập các sơ đồ quan hệ được sử dụng để biểu diễn thông tin đến đối với 1 tổ chức

Khoá

- Đ/n
  - Cho $R(A_1, A_2, \ldots A_n)$, $K \subseteq \{A_i\}$,
  - K là khoá nếu với $\forall t_1, t_2 \in R, \exists A_i \in K: t_1.A_i \neq t_2.A_i$

SINH_VIEN (maSV,tenSV,ngaysinh,nam, diachi,lop)

- T/c:
  - $K \subseteq K', K' \subseteq \{A_i\}$, K là khoá $\Rightarrow$ K' cũng là khoá

SINH_VIEN (maSV,tenSV,ngaysinh, nam, diachi,lop)
SINH_VIEN (maSV,tenSV,ngaysinh, nam, diachi,lop)

Phân loại khoá

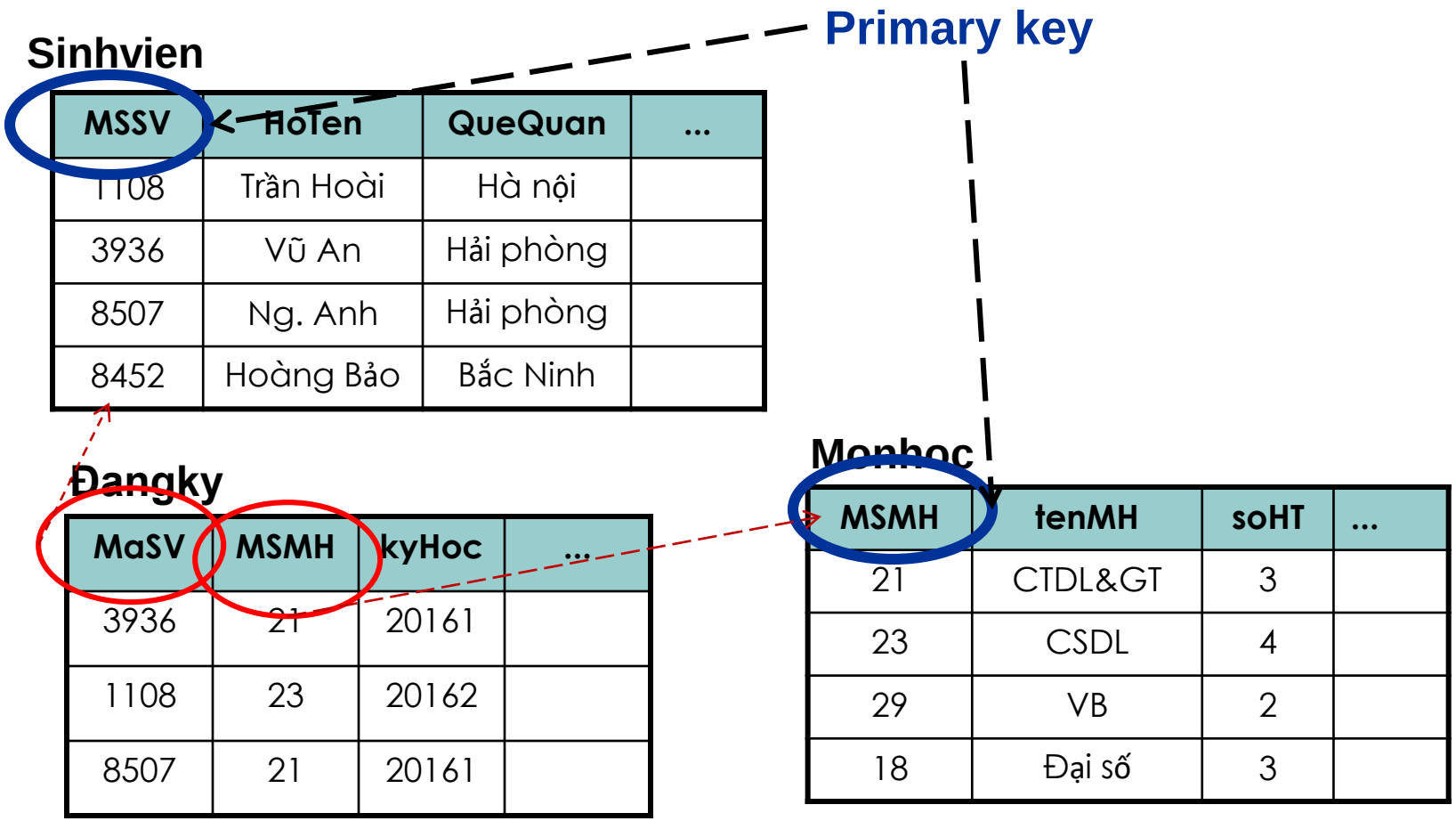
- Khoá tối thiểu
  - Cho $R(A_1,A_2, \dots A_n)$, $K \subseteq \{A_i\}$,
  - K là khoá tối thiểu nếu K là khoá và $!\exists K' \subset K$ mà K' là khoá
  - 1 khóa tối thiểu tốt được chọn là khóa chính

- Khoá ngoài
  - Cho $R(A_1,A_2, \dots A_n)$, $K \subseteq \{A_i\}$, $R'(A'_1,A'_2, \dots A'_m)$
  - K là khoá ngoài của R' tham chiếu đến quan hệ R nếu K là khoá chính của R

SINH_VIEN (**maSV**,tenSV,ngaysinh, nam, diachi,***malop***)

LOP(**malop**,lop, khoa, GVCN, loptruong)

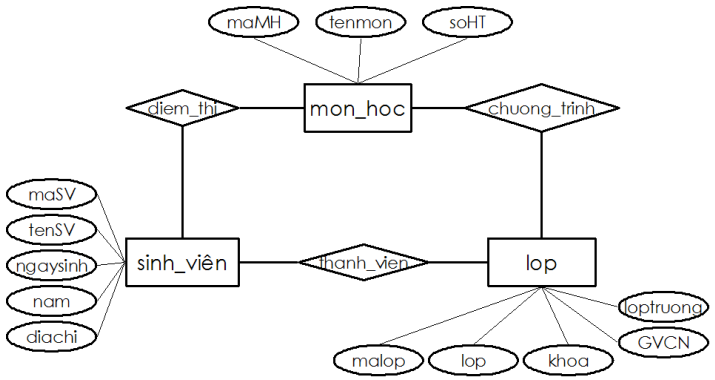
Primary key vs Foreign key



Các bước xây dựng một CSDL

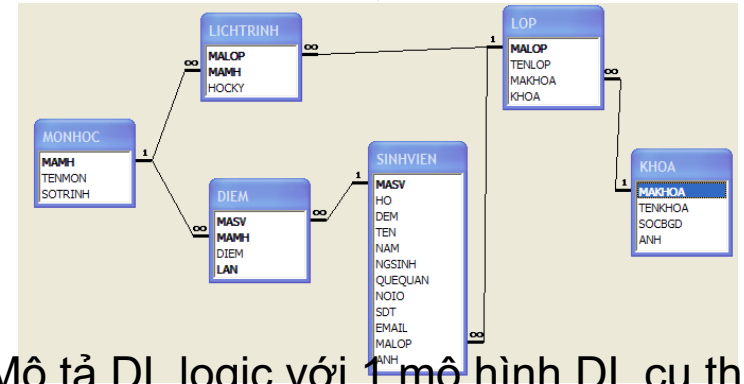
Mô tả ứng dụng

1: PHÂN TÍCH



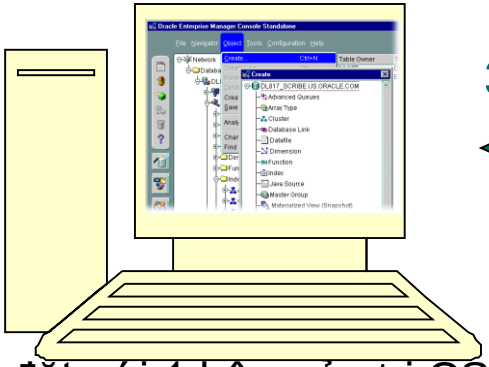
Mô hình hoá DL (vd: Sơ đồ thực thể-liên kết)

2: THIẾT KẾ



Mô tả DL logic với 1 mô hình DL cụ thể
(vd: Sơ đồ quan hệ)

3: CÀI ĐẶT



Cài đặt với 1 hệ quản trị CSDL
(vd: ORACLE)

Kết luận

- Điểm khác nhau của các mô hình dữ liệu
  - khả năng biểu diễn dữ liệu về mặt ngữ nghĩa
  - khả năng biểu diễn truy vấn dữ liệu
  - hiểu quả của cài đặt trong máy tính
- Mô hình thực thể-liên kết cho phép biểu diễn dữ liệu gần với nhận thức của con người
- Mô hình quan hệ
  - Cho phép biểu diễn logic dữ liệu
  - dễ ánh xạ sang cấu trúc lưu trữ vật lý
  - Dựa trên nền tảng toán học cho phép tối ưu hoá các truy xuất dữ liệu

"More than 90% of current database applications are built on relational database systems which utilise relational model as its underlying data model"*

* R. Elmasri and S. Navathe. Fundamentals of Database Systems

Các điểm cần lưu ý

- Khái niệm chung về mô hình dữ liệu
- Các mô hình dữ liệu
- Mô hình thực thể - liên kết
- Mô hình dữ liệu quan hệ
- Các bước xây dựng một CSDL
- Biến đổi: sơ đồ thực thể - liên kết ➔ Sơ đồ quan hệ

Tài liệu tham khảo

[Cattel,1997] R. G. G Cattel et al. The Object Database Standard: ODMG 2.0. *Morgan-Kaufmann*, 1997

[Chen, 1976] The entity-relationship model: toward a unified view of data. TODS, 1(1), 1976

[Chen, 2002] P. Chen. Entity-Relationship Modeling: Historical Events, Future Trends, and Lessons Learned. *Software Engeenering*, 2002

[Codd, 1980] E. F. Codd. Data Models in Database Management. *Proc. of SIGMOD*, 1980

[Codd, 1970] E. F. Codd. The relational model of data for large shared data banks. *CACM*, 13(6), 1970

[McHugh,1997] J. McHugh et al. Lore: A Database Management System for Semistructured Data. *SIGMOD Record*, 26(3), 1997

[XML] Extensible Markup Language. http://www.w3.org/XML

[Zdonik,1986] S. B. Zdonik et al. Language and Methodology for Object-Oriented Database Environments. *Proc. Of the Hawai International Conference on System Sciences*, 1986