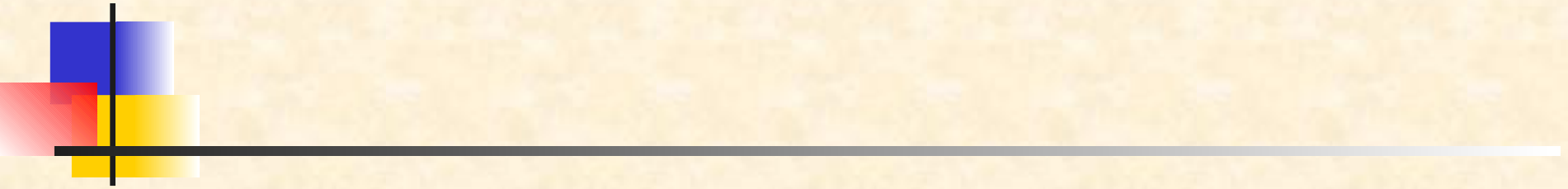
Một số khái niệm đầu tiên về các hệ thống cơ sở dữ liệu



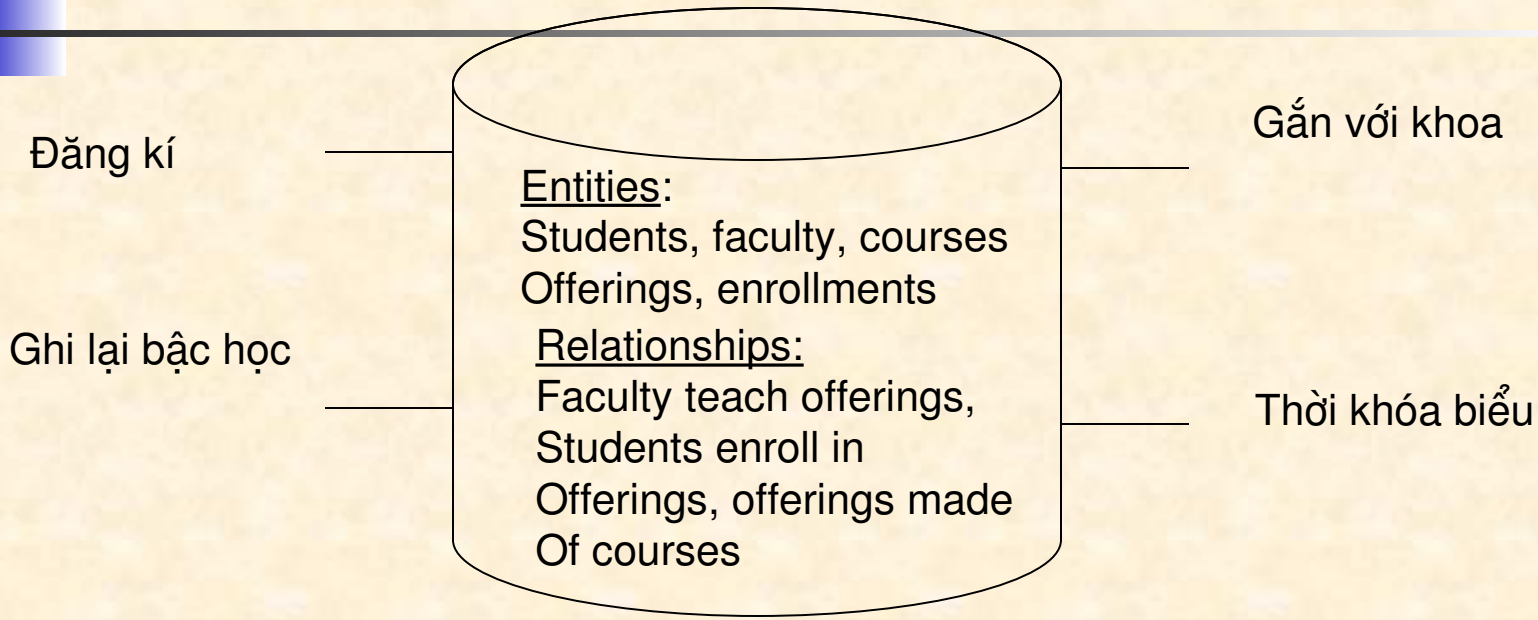
Cơ sở dữ liệu

- Định nghĩa
  - Là tập các dữ liệu bền vững, có thể chia sẻ.
- Lí do cần thiết dùng cơ sở dữ liệu
  - Công việc xử lí dữ liệu hàng ngày, thông tin về sách, về ngân hàng, đặt hàng...
  - Dữ liệu thể hiện các sự kiện thường ngày
  - Cần thiết tổ chức dữ liệu để truy cập dễ dàng

Đặc tính của cơ sở dữ liệu

- **Bền vững** – tức dữ liệu được đặt trên thiết bị lưu trữ ổn định, cho phép sử dụng nhiều lần
- **Chia sẻ** – tức cơ sở dữ liệu cho phép nhiều người dùng, nhiều công việc.
  - Cơ sở dữ liệu cá nhân
  - Cơ sở dữ liệu nhóm
  - Cơ sở dữ liệu xí nghiệp
- **Liên kết** – tức dữ liệu được lưu tại nhiều nơi, có liên kết, như bức tranh tổng thể

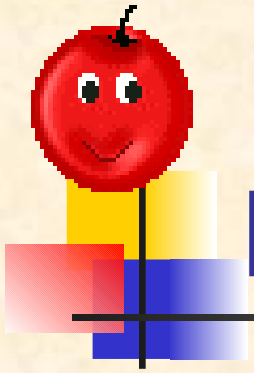
Thí dụ cơ sở dữ liệu về đại học



Cho phép biết người đứng lớp

Cho biết tên sinh viên, với lớp học

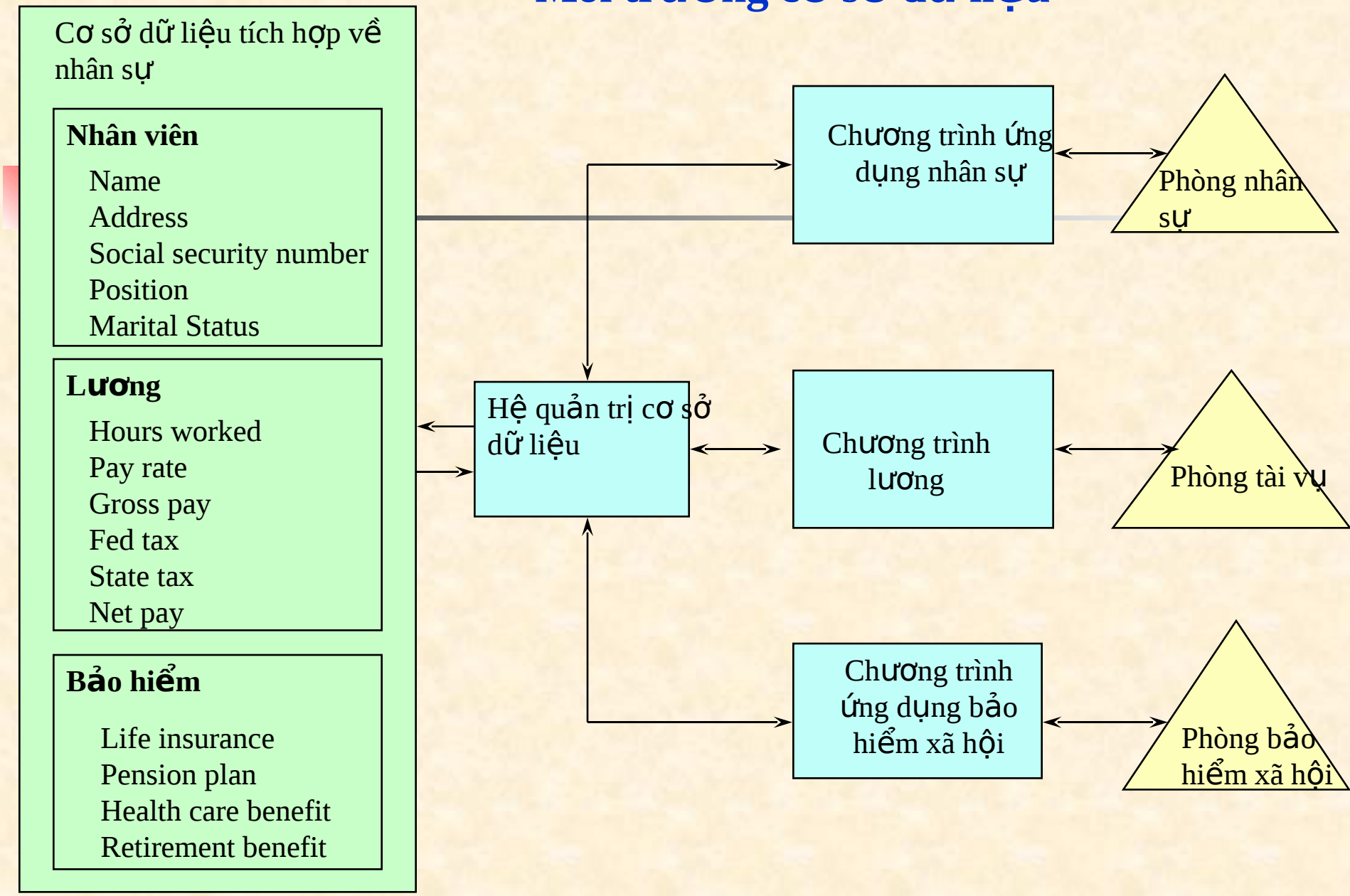
Biết được khoa, bậc học



Hệ quản trị cơ sở dữ liệu

- Hệ quản trị cơ sở dữ liệu (database management system - DBMS) là tập các phần mềm cho phép tạo, sử dụng, bảo trì cơ sở dữ liệu
- Trước tiên, DBMSs đảm bảo lưu trữ và tìm kiếm hiệu quả
- Hệ quản trị là phần mềm thương mại
- Theo thị trường, hệ quản trị có các khía cạnh : thu thập dữ liệu, lưu trữ, bảo trì, lập báo cáo...

Môi trường cơ sở dữ liệu



Khía cạnh thương mại của hệ quản trị cơ sở dữ liệu

- **Xác định cơ sở dữ liệu** – các công cụ ngôn ngữ và đồ họa cho phép xác định thực thể, mối quan hệ, điều kiện ràng buộc, quyền truy cập...
- **Truy cập phi thủ tục** – các công cụ ngôn ngữ và đồ họa cho phép truy cập dữ liệu mà không phải viết chương trình phức tạp
- **Phát triển ứng dụng** – công cụ đồ họa để phát triển thực đơn, khuôn dạng nhập dữ liệu, báo cáo...
- **Giao diện ngôn ngữ phi thủ tục** – ngôn ngữ kết hợp truy cập phi thủ tục với các khả năng của ngôn ngữ lập trình đầy đủ
- **Xử lí giao tác** – cơ chế điều khiển để tránh xung đột dữ liệu và khôi phục sai sót
- **Tinh chỉnh dữ liệu** – công cụ giám sát và nâng cao hiệu năng hệ thống

Hai khung nhìn cơ sở dữ liệu

- **Khung nhìn vật lí** : mô tả nơi lưu dữ liệu
  - Thiết bị, đĩa, rãnh, bề mặt, từ quạt, bản ghi...
  - Băng từ, khối dữ liệu, số các bản ghi...
- **Khung nhìn logic**: mô tả ứng dụng cần đến dữ liệu
  - Sự kiện cần thiết của xí nghiệp
  - Tên, độ dài bản ghi, kiểu dữ liệu

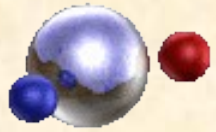
- DBMS cho phép người dùng hay người lập trình không phải quan tâm đến nơi, cách thức lưu trữ dữ liệu

Tiến hóa của công nghệ cơ sở dữ liệu

Kỉ nguyên	Thế hệ	Định hướng	Nét chính
1960s	thế hệ 1	File	cấu trúc file và giao diện chương trình
1970s	thế hệ 2	mạng	mạng, phân cấp các bản ghi chương trình chuẩn
1980s		quan hệ	quan hệ tối ưu, xử lí giao tác
1990s	đối tượng		đa phương tiện cơ sở dữ liệu động, xử lí phân tán

Chia sẻ thị trường về hệ quản trị cơ sở dữ liệu xí nghiệp

Sản phẩm	Chiếm thị trường	Bình luận
IBM DB2	33%	Dominates the MVS and AS/400 environments
Oracle	29%	Dominates the Unix environment (61%), leader in NT environment (46%)
Microsoft SQL server	10%	30% market share in NT environment; no presence in other environments
Informix	4%	13% share of Unix market
Sybase SQL server	4%	
Khác	20%	Includes CA, NCR, Progress Software, NEC etc.



Xác định cơ sở dữ liệu

- Để xác định cơ sở dữ liệu, cần xác định thực thể và mối quan hệ
- DBMS dùng các bảng để lưu tập các thực thể. Mối quan hệ nhằm vào các liên kết giữa các bảng
- Ngôn ngữ mô tả dữ liệu (DDL) xác định mỗi phần tử dữ liệu như là bản ghi trong bảng, trước khi phần tử dữ liệu được chuyển sang dạng dùng cho người lập trình
- Ngôn ngữ xử lí dữ liệu (DML) thông dụng là SQL

Người dùng hay dùng SQL, *Structured Query Language*

- Cho phép tìm kiếm phức tạp, với điều kiện

  SELECT tên FROM sinh viên WHERE toán > 7 and tuổi < 30

Ngôn ngữ khác : QUEL, QBE (Query by Example, QBE)

Ngôn ngữ SQL

Tổ chức dữ liệu

- 4 tiếp cận
  - Mô hình cơ sở dữ liệu quan hệ
  - Mô hình cơ sở dữ liệu phân cấp
  - Mô hình cơ sở dữ liệu mạng

  - Mô hình cơ sở dữ liệu đa chiều

Mô hình cơ sở dữ liệu quan hệ

- Sử dụng bảng 2 chiều
- Bảng được gọi là quan hệ
- Dựa trên lí thuyết tập
- Dòng dữ liệu = bản ghi = bộ (tuple)
- Cột dữ liệu = trường = thuộc tính
- Dùng tập các bảng thay vì một bảng, để tạo nên cơ sở dữ liệu

Thí dụ mô hình quan hệ

East Coast Mgrs.

NAME	TITLE	AGE	DIVISION
Smith, A.	Dir., Accntg	43	China
Jones, W.	Dir., TQM	32	Stemware

Thực thể, thuộc tính, giá trị

- Thực thể : là người, đối tượng hay sự kiện mà người ta đang quản lí thông tin về chúng
- Thuộc tính : tính chất, chất lượng, mô tả thực thể cụ thể
- Trường khóa : thông tin xác định duy nhất bản ghi, để có thể tìm kiếm, lưu trữ và sắp xếp

Các bước thiết kế cơ sở dữ liệu

Yêu cầu của người dùng

Thiết kế khái niệm

(có thể là mô hình thực thể quan hệ)

Thiết kế logic

(có thể mô hình quan hệ)

thiết kế vật lí

(tối ưu về hiệu năng)

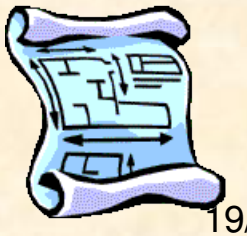
- **yêu cầu cơ sở dữ liệu từ phía người dùng**

- **Khung nhìn người dùng : là phần cơ sở dữ liệu, là quan trọng đối với người dùng**

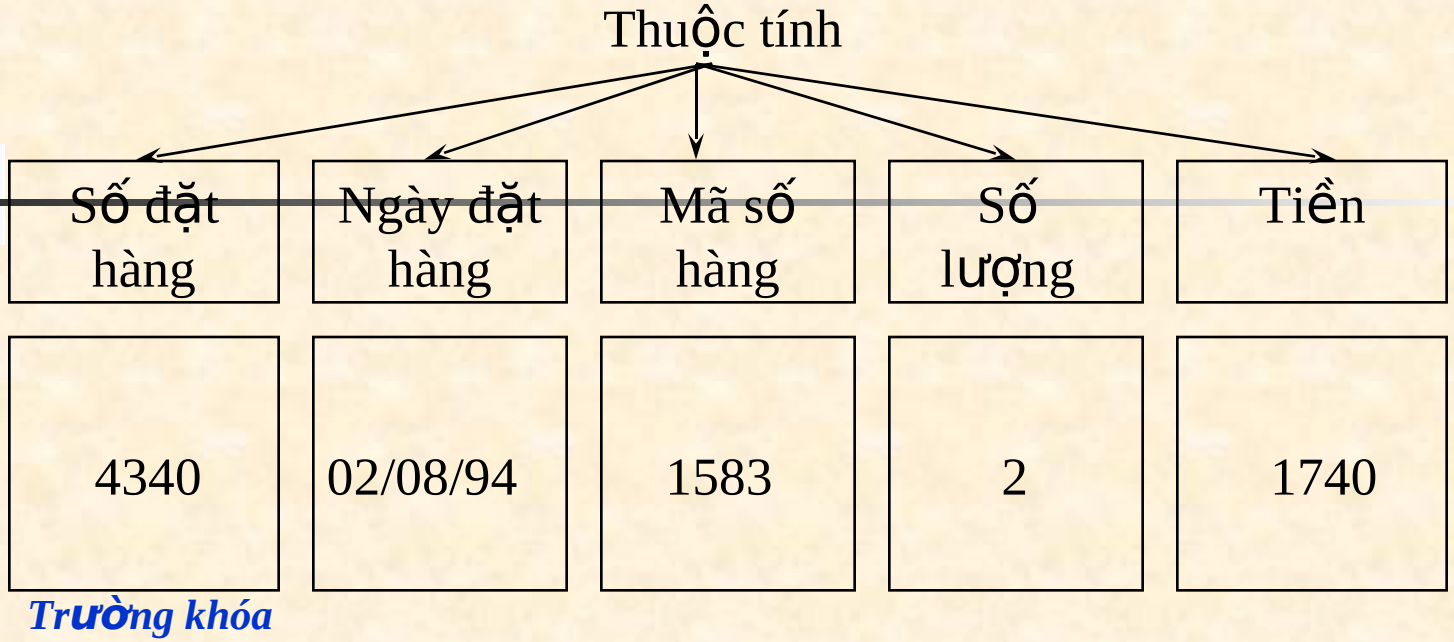
- **Một số khó khăn**

Thiết kế khái niệm
Mô hình thực thể quan hệ

- Thực thể
  - Là cái đang quan tâm
    - Có thể là trực tiếp, hay gián tiếp.
- Thuộc tính – tính chất, điều vốn có của thực thể
- Khóa
  - Mỗi thực thể cần được xác định duy nhất bằng thuộc tính, hay nhóm các thuộc tính, gọi là khóa



Thực thể Đặt hàng



Bản ghi này mô tả thực thể ORDER và các thuộc tính. Các giá trị riêng đối với yêu cầu đặt hàng là các giá trị thuộc tính. Trường khóa là *Order number* do mỗi đặt hàng gắn với con số duy nhất.

Order : [Order number , order date, item number, quantity, amount]

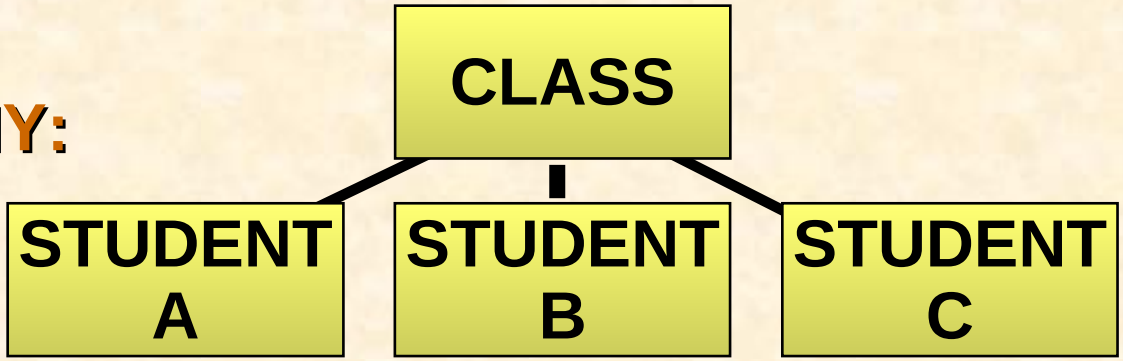
Mối quan hệ

- Là liên kết giữa các thực thể.
  - Có dạng 1 ngôi (quản lí), 2 ngôi, nhiều ngôi
- Bậc quan hệ chỉ số của mỗi thực thể tham gia trong mối quan hệ :
  - Một một (1:1)
  - Một nhiều (1:N)
  - Nhiều nhiều (N:M)
- Mối quan hệ M:N có thể có thuộc tính

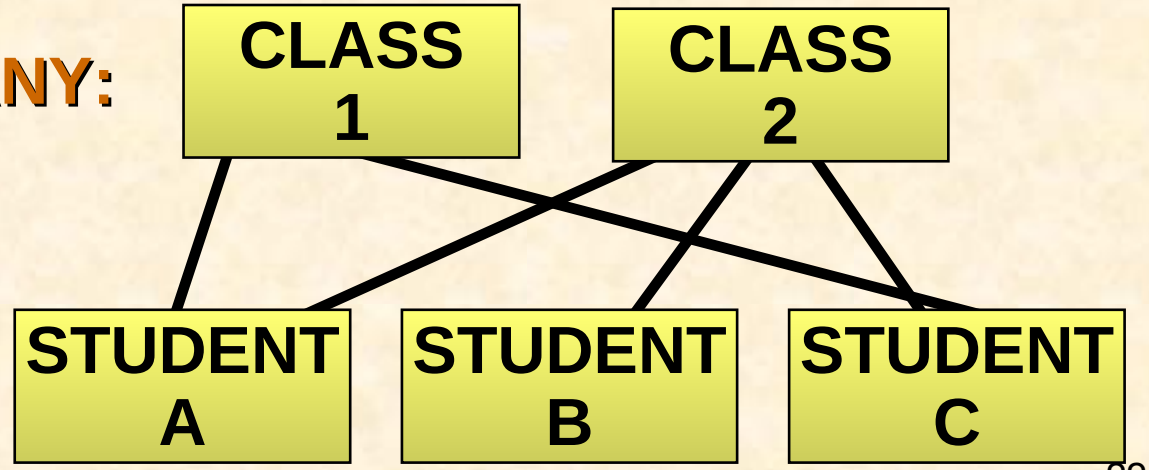
Kiểu quan hệ

ONE-TO-ONE:

ONE-TO-MANY:



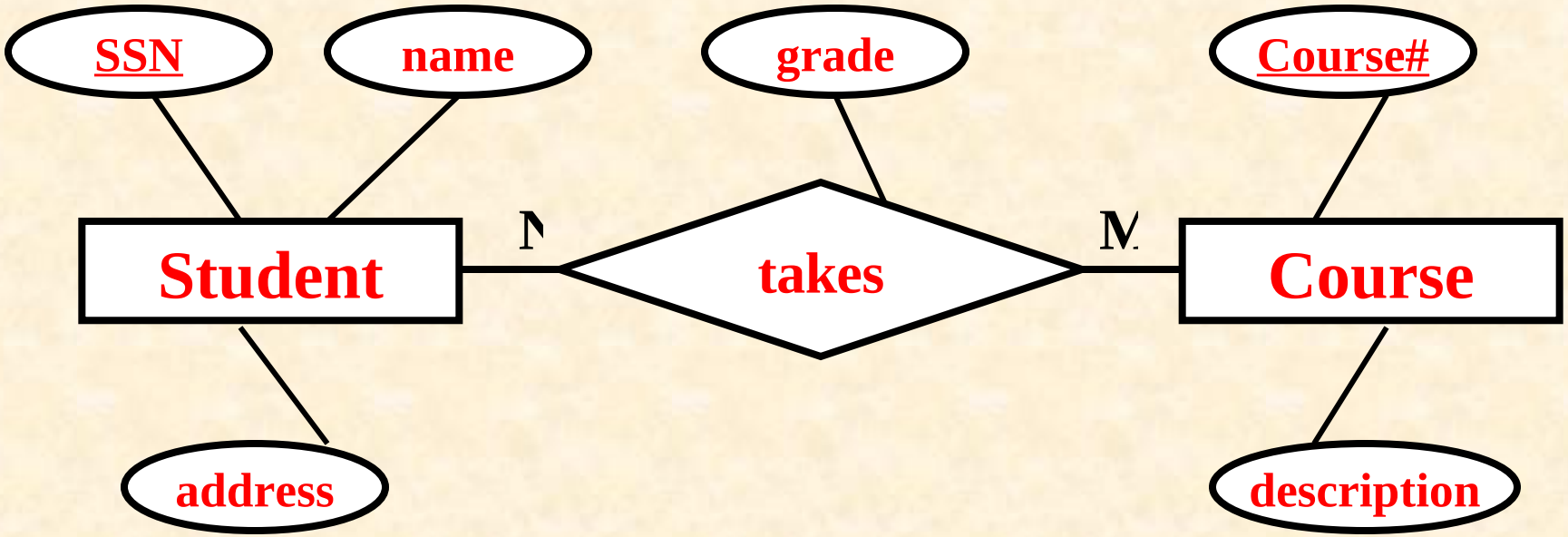
MANY-TO-MANY:



Sơ đồ thực thể quan hệ

- Mỗi thực thể được thể hiện bằng hình chữ nhật.
- Mối quan hệ được thể hiện qua hình thoi.
- Các thuộc tính được thể hiện qua hình elip.
- Các bậc viết kề bên thực thể

Sơ đồ E-R (diagram)



Chuyển mô hình ER sang mô hình quan hệ

Dùng quan hệ để thể hiện thực thể

Quan hệ N:M được thể hiện bằng quan hệ tách biệt

Khóa là nối kết các khóa thực thể.

Các thuộc tính của mối quan hệ không là khóa.

Khóa ngoài được xác định “thuộc tính của một bảng quan hệ, là khóa chính của quan hệ khác”

Thí dụ Student-Course

Thiết kế cơ sở dữ liệu để theo dõi sinh viên với bài giảng, và kết quả học nhận được

Thực thể : Student: [SSN, name, address]
Course: [Course-Id, description]

Mối quan hệ Student takes Course: [grade]
N : M

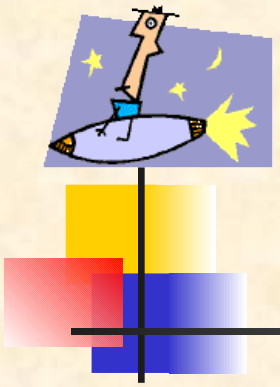
Thí dụ về SQL

- Tìm tên và số bảo hiểm của sinh viên.

```
SELECT Name, SSN
FROM Student;
```

Name	SSN
M. Tomkins	111-22-3333
L. Richardo	444-71-2222
H. McEnroe	795-44-1111

```
SELECT  cột
FROM    bảng
WHERE   điều kiện
```



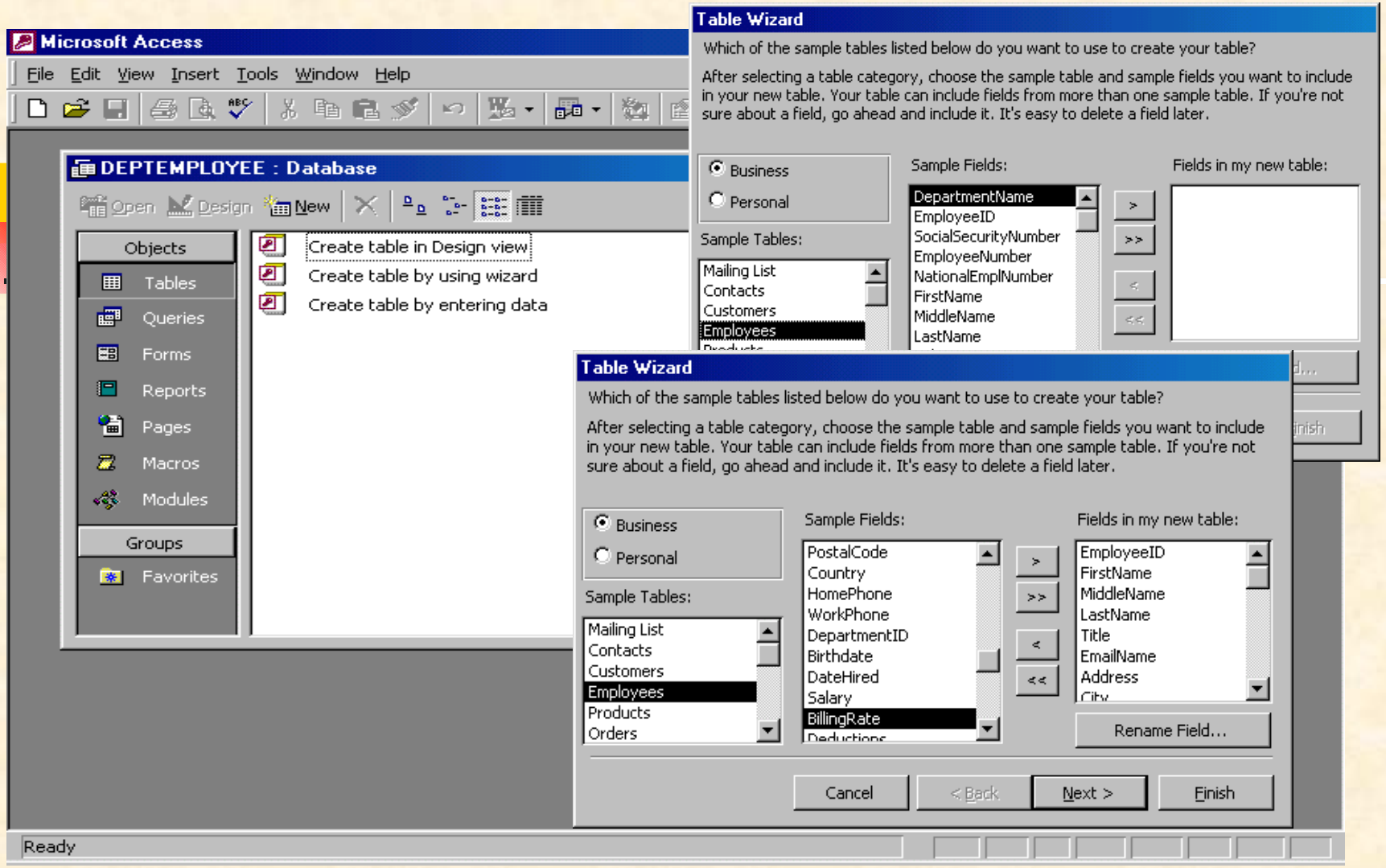
Điều kiện và nối trong SQL

- Tìm SSN của sinh viên đạt kết quả B.

```
SELECT  SSN
FROM    Takes
WHERE  Grade = B;
```

- Có thể nối hai bảng và rút ra thông tin. Thí dụ tìm tên các sinh viên kết quả A.

```
SELECT  Student.Name
FROM     Student, Takes
WHERE   Takes.grade = A AND Takes.SSN = Student.SSN
```



MS Access cho phép thể hiện thực thể như các bảng. Trong bảng có thuộc tính. Người ta có thể tạo bảng theo nhiều cách, như trong hình, theo Wizard...

Lí do cần mô hình hóa dữ liệu

- Một cơ sở dữ liệu cần thể hiện thế giới thực
- Chỉ mô hình hóa mới thể hiện được thế giới thực
- Mô hình hóa nhấn mạnh thể hiện thực tế, sự phức tạp của kinh doanh
- Thể hiện đồ họa tốt cho thực tế và cả dữ liệu trong cơ sở dữ liệu
- Đích của công việc là định tên sự kiện trong cơ sở dữ liệu

Phát triển hệ thống thông tin dựa trên dữ liệu

Lĩnh vực bài toán → Thiết kế khái niệm

↓

Conceptual Schema, e.g., ER Model

↓

thiết kế logic

↓

Logical Schema, e.g., Relational Model

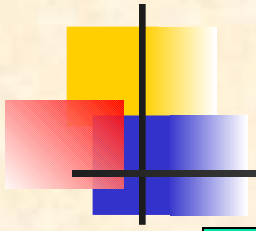
↓

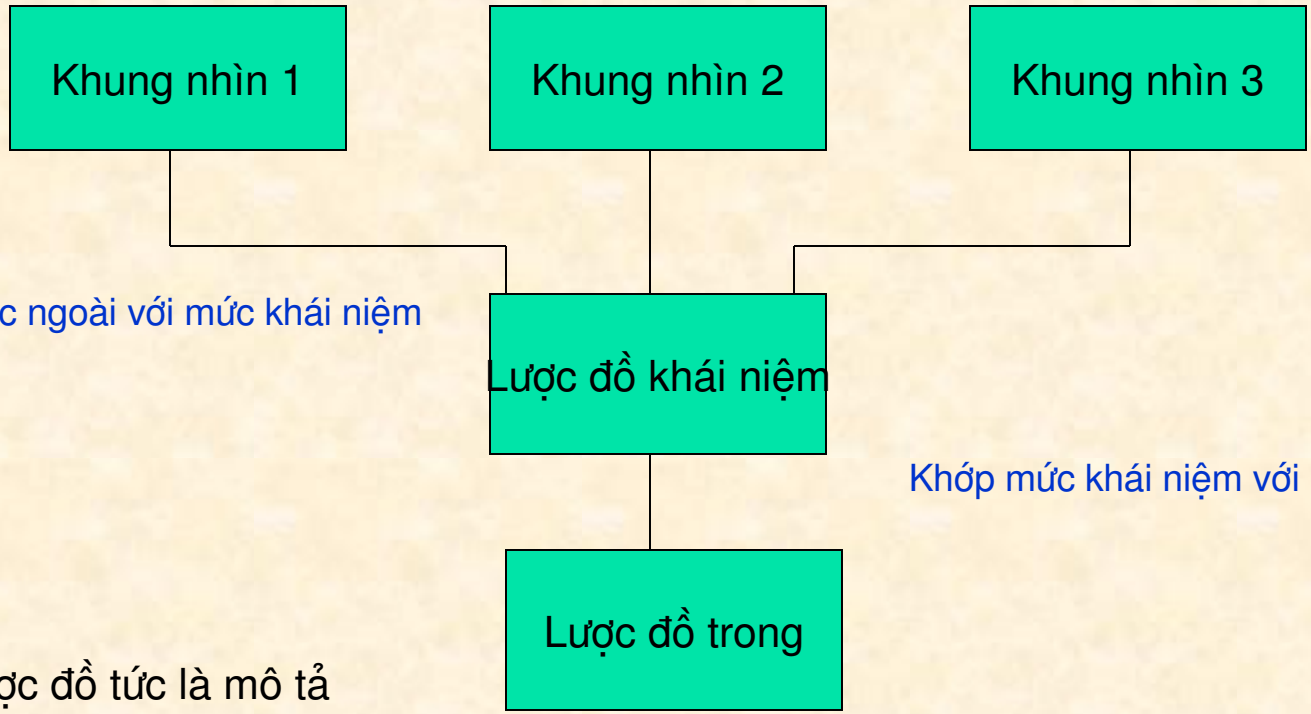
thiết kế vật lí

↓

Physical Schema, e.g., Access implementation

Ba kiến trúc lược đồ





Độc lập dữ liệu có nghĩa cơ sở dữ liệu cần được xác định độc lập với chương trình ứng dụng mà nó sử dụng, như báo cáo, khuôn dạng...

Quản lí dữ liệu

vấn đề kinh doanh thế giới thực

- Một số khó khăn
  - **Khối lượng dữ liệu tăng nhanh, mà cần giữ dữ liệu**
  - **Nhiều truy cập, sử dụng các nguồn, thiết bị khác nhau**
  - **Chỉ phần nhỏ dữ liệu trong tổ chức được dùng trong trợ giúp quyết định**
  - **Dữ liệu ngoài cũng cần thiết cho quyết định**

Quản lí dữ liệu …

- Dữ liệu thô có trong nhiều hệ thống hợp pháp
- Yêu cầu mang tính pháp luật đối với dữ liệu khác nhau ở các nước
- Có nhiều công cụ quản trị dữ liệu
- Cần có an toàn, toàn vẹn dữ liệu

- Nhận xét chung, dữ liệu cần :
  - Có tính thời sự
  - Chính xác

Quản lí dữ liệu …

- Về lịch sử, dữ liệu được tổ chức phân cấp để quản lí các giao tác
- Phân cấp là hiệu quả đối với xử lí tác nghiệp, số lượng lớn các dữ liệu
- Mô hình trước (mạng, phân cấp) không tiện cho quản trị, cho hỏi dữ liệu
- Cơ sở dữ liệu quan hệ có nhiều chức năng

Quản lí dữ liệu …

- Cơ sở dữ liệu quan hệ
  - Tiện lợi cho tính toán người dùng qua định nghĩa đơn giản, các câu hỏi, khuôn dạng, báo cáo...
  - Có trợ giúp quyết định
- Đối với kiến trúc khách/ chủ, cơ sở dữ liệu trở nên phân tán
- Cơ sở dữ liệu nhiều chiều và nhiều khối dữ liệu cần đến kiến thức về kho dữ liệu

Chu kì vòng đời dữ liệu và phát hiện tri thức

- Y tưởng chính:
  - Chuyển hóa dữ liệu, thông tin và tri thức
  - Quá trình :
    - Thu thập dữ liệu từ nhiều nguồn
    - Lưu trong cơ sở dữ liệu
    - Làm sạch dữ liệu và lưu trong kho dữ liệu
    - Thực hiện khai phá dữ liệu, thu được tri thức
    - Lưu kết quả trong cơ sở tri thức

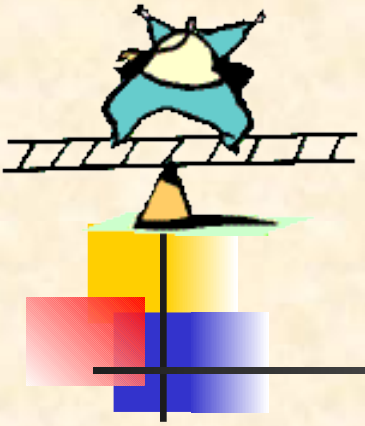
Chuyển dữ liệu sang tri thức

Nguồn dữ liệu và thu thập dữ liệu

- Dữ liệu gồm
  - Tư liệu
  - Hình ảnh
  - Bản đồ
  - Âm thanh
  - Hoạt hình
  - Khái niệm, ý nghĩ, ý kiến
- Dữ liệu có thể là thô, hay đã được tổng hợp
- Dữ liệu có thể là từ trong, bên ngoài hay của cá nhân

Kiểu dữ liệu

- Dữ liệu trong
  - Con người, sản phẩm, quá trình,
- Cá nhân
  - Các đánh giá chủ đề, ý kiến, qui tắc. Một số dữ liệu là “ẩn”, số khác là hiện.
- Dữ liệu ngoài
  - Các báo cáo, cơ sở dữ liệu ngoài, hình ảnh



Thu thập dữ liệu

- Vấn đề :
  - Chất lượng dữ liệu
  - Toàn vẹn dữ liệu, tức thay đổi dữ liệu tại một nơi sẽ truyền khắp các nơi
- Phương pháp thu thập
  - Thủ công
  - Có công cụ, đầu dò, thu
  - Quét hay tải tự động

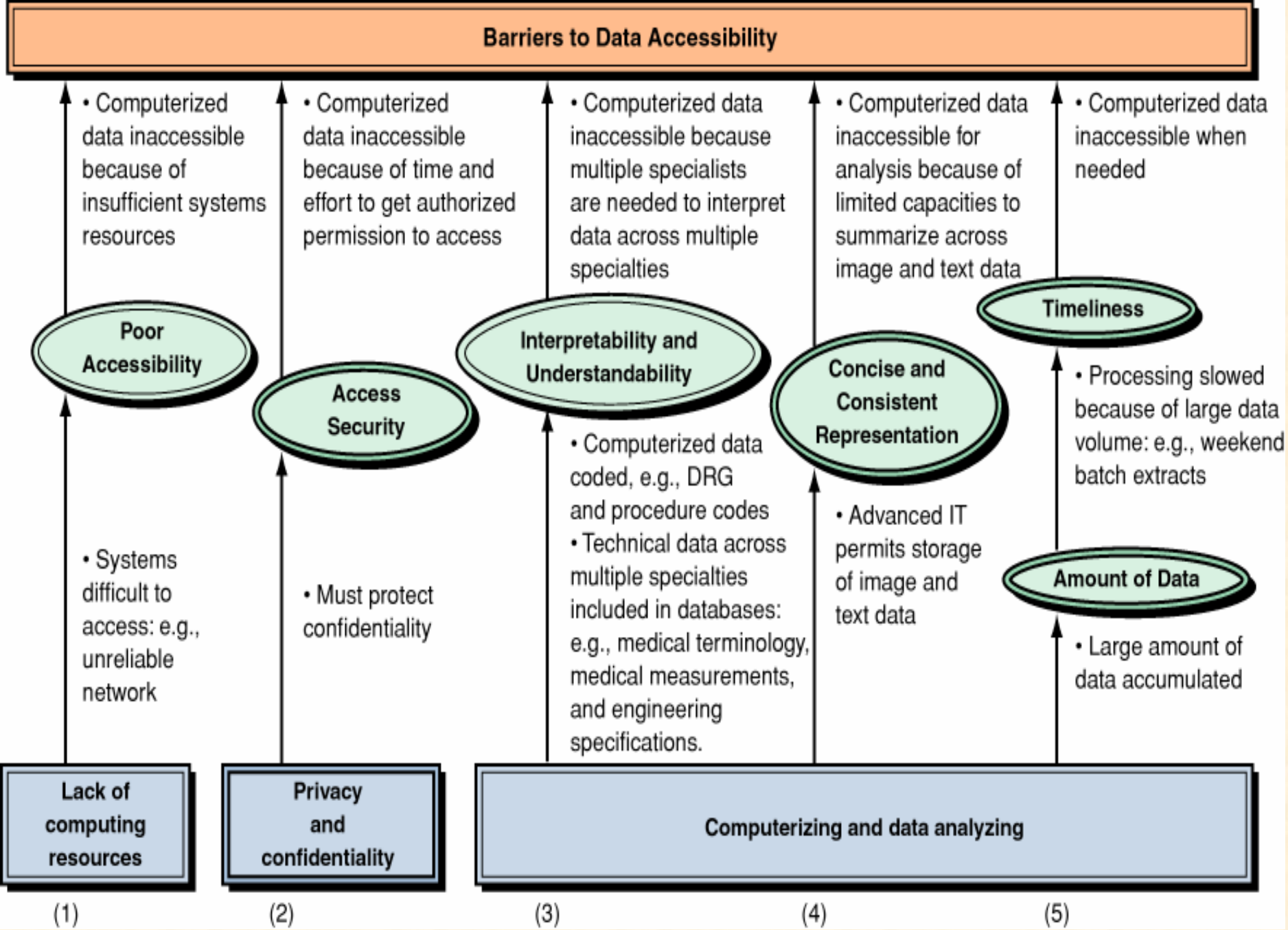
- Quản lí luồng dữ liệu (Data flow manager DFM)
  - Trợ giúp thu thập dữ liệu từ nhiều nguồn
  - Có hệ thống DSS, bộ xử lí dữ liệu trung tâm, bộ toàn vẹn dữ liệu, nối với nguồn ngoài

Chất lượng dữ liệu

- Chất lượng dữ liệu quyết định tính sử dụng được
- Các sai sót tiềm năng
  - Dữ liệu không chính xác
  - Dữ liệu mơ hồ, mờ
  - Dữ liệu không được chỉ số hóa đúng
  - Không có dữ liệu đang cần

Tham số về chất lượng dữ liệu :

Về tính chất	Chính xác, đối tượng, tin cậy, uy tín
Về truy cập	Truy cập được, an toàn truy cập
Về ngữ cảnh	Thể hiện, có giá trị, đầy đủ
Về thể hiện	Dễ hiểu, diễn tả, chi tiết



o cản truy cập dữ liệu

Xử lí phân tích so với xử lí giao tác

- Xử lí phân tích yêu cầu xử lí hàng ngày các giao tác về tổ chức như thanh taón, đặt hàng...
- Các cơ sở dữ liệu và hệ thống xử lí cần thiết được gọi là hệ thống tác nghiệp
- Nhu cầu thị trường đòi hỏi có xử lí phân tích

- Xử lí phân tích trực tuyến (OLAP) cho phép người dùng
  - Truy cập dữ liệu dễ dàng
  - Ra quyết định nhanh
  - Thao tác chính xác và hiệu quả
  - Mềm dẻo
- OLAP thường đi với DSS, EIS và các hoạt động hướng người dùng khác.

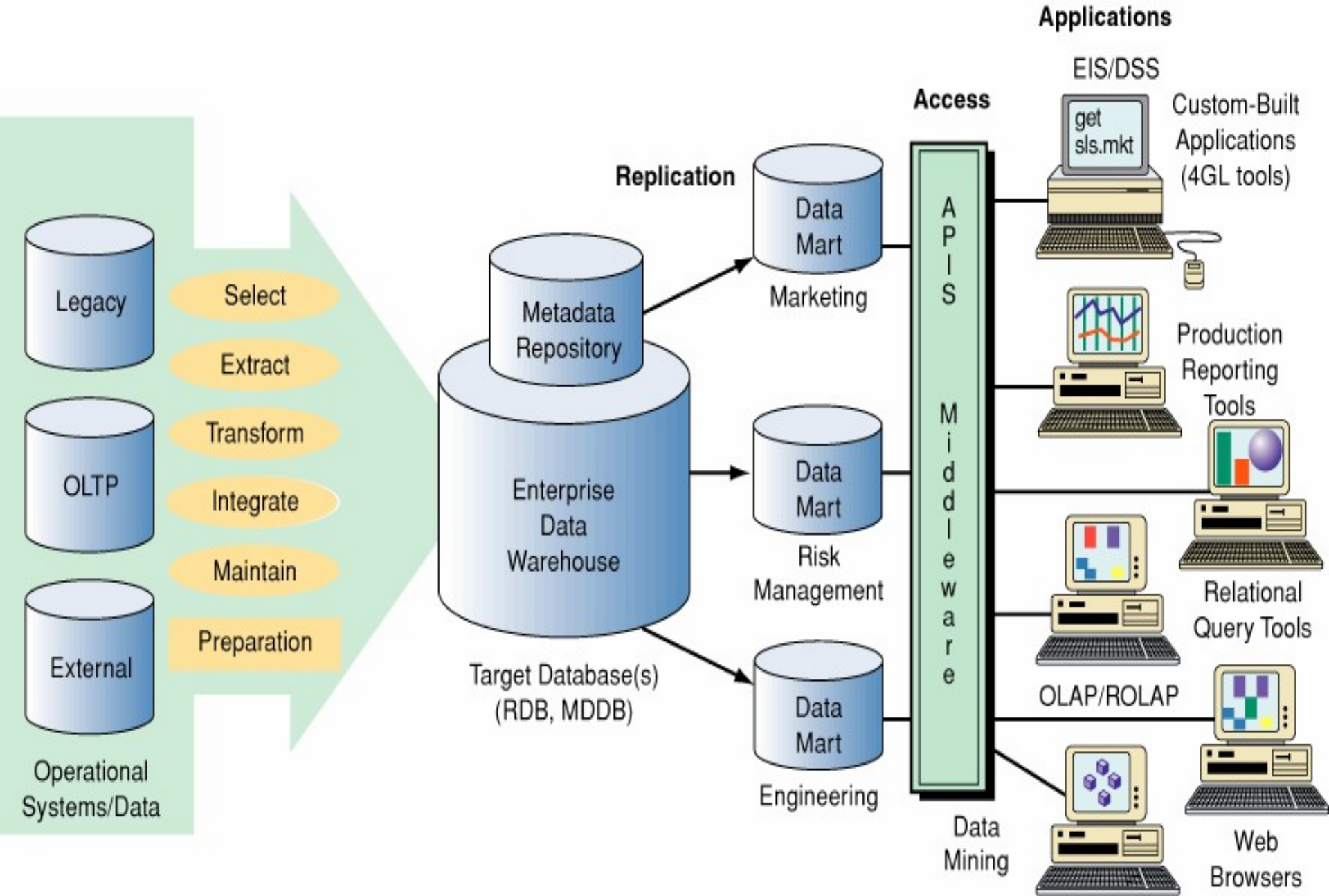
Xử lí phân tích

- OLAP cần đến ba khái niệm:
- **Thể hiện kinh doanh về dữ liệu người dùng**
- **Môi trường khách/ chủ cho phép người dùng hỏi về dữ liệu và ra báo cáo**
- **Thư mục trên máy tính chủ, kho dữ liệu... Cho phép an toàn dữ liệu và điều khiển tập trung dữ liệu**

Kho dữ liệu

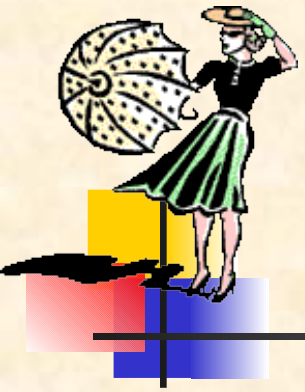
- Định nghĩa:
  - Tập dữ liệu hướng chủ đề, tích hợp, thay đổi theo thời gian, không mất đi, để ra quyết định
- Xử lí phân tích trực tuyến (OLAP)
- Khai phá dữ liệu
- Tiếp thị cơ sở dữ liệu
- Khớp nhu cầu về dữ liệu và nhu cầu h
  thống thông tin điều hành

Kho dữ liệu


Applications
EIS/DSS
Access
Custom-Built Applications (4GL tools)
get sls.mkt
Replication
Data Mart
Marketing
Metadata Repository
Legacy
Select
Extract
Transform
Integrate
Maintain
Preparation
OLTP
External
Enterprise Data Warehouse
Target Database(s) (RDB, MDDB)
Data Mart
Risk Management
Data Mart
Engineering
Operational Systems/Data
API
Middleware
Production Reporting Tools
Relational Query Tools
OLAP/ROLAP
Data Mining
Web Browsers

Các tính chất

- Cơ sở dữ liệu tách biệt
- Dữ liệu sạch
- Thông tin tổng hợp để quản lí
- Dữ liệu meta
- Dữ liệu từ hệ thống hợp pháp
- Thời gian là yếu tố dữ liệu



Khai phá dữ liệu

- Kiểu dữ liệu
  - Liên kết
  - Xâu, chuỗi
  - Phân loại, tức các luật
  - Phân cụm, tức tạo nhóm
  - Dự báo, theo chuỗi thời gian
- Phát hiện giả thuyết, thay vì thử giả thuyết

Cám ơn sự theo dõi