CHƯƠNG 1.
KHÁI NIỆM CƠ BẢN VỀ CƠ SỞ DỮ LIỆU

(Phần 1)

Nguyễn Đình Hóa
dinhhoa@gmail.com / hoand@ptit.edu.vn
094-280-7711

KHÁI NIỆM CƠ BẢN VỀ CSDL

- Các khái niệm cơ bản: Cơ sở dữ liệu, Hệ quản trị CSDL, Hệ cơ sở dữ liệu
- Các hệ CSDL truyền thống
- Các thành phần của một hệ quản trị CSDL
- Sự cần thiết của việc thiết kế CSDL
- Các vai trò trong môi trường CSDL
- Mô hình trừu tượng 3 lớp
- Các ngôn ngữ cơ sở dữ liệu
- Phân loại các hệ CSDL

KHÁI NIỆM VỀ CSDL

- Theo nhận thức chung nhất, ***cơ sở dữ liệu*** (database) đơn giản là một tập thông tin (dữ liệu) có liên quan đến nhau.

  => định nghĩa này rất mơ hồ, vì theo đây, có thể coi một trang văn bản là một CSDL.

- Khái niệm ***"dữ liệu"*** trong CSDL có thể bao gồm một phạm vi rất rộng các đối tượng: chữ số, văn bản, đồ họa, video,...

- Định nghĩa cụ thể hơn của một CSDL bao gồm một tập các đặc tính không tường minh được xem xét cùng nhau để định nghĩa một CSDL.

KHÁI NIỆM VỀ CSDL (cont.)

- CSDL thể hiện các khía cạnh khác nhau của thế giới thực.
- CSDL được coi là một tập dữ liệu gắn kết logic với nhau. Các dữ liệu ngẫu nhiên không được coi là một CSDL (mặc dù chúng là những ngoại lệ).
- Một CSDL được thiết kế, xây dựng và sử dụng cho một số mục đích cụ thể. Nó được sử dụng bởi một tập người dùng và ứng dụng cụ thể ngay từ khi mới thiết kế.

HỆ QUẢN TRỊ CSDL

- ***Hệ quản trị CSDL*** (DBMS – Database management system) là một hệ thống phần mềm cho phép tạo lập CSDL và điều khiển mọi truy nhập đến CSDL đó.
- Các đặc tính quan trọng của một hệ quản trị CSDL:

1. Cho phép người dùng tạo mới CSDL, thông qua ngôn ngữ định nghĩa dữ liệu (DDLs – Data Definition Languages).
2. Cho phép người dùng truy vấn cơ sở dữ liệu, thông qua ngôn ngữ thao tác dữ liệu (DMLs – Data Manipulation Languages).
3. Hỗ trợ lưu trữ số lượng lớn dữ liệu, thường lên tới hàng Gigabytes hoặc nhiều hơn, trong một thời gian dài. Duy trì tính bảo mật và tính toàn vẹn trong quá trình xử lý.
4. Kiểm soát truy nhập dữ liệu từ nhiều người dùng tại cùng một thời điểm.

HỆ CƠ SỞ DỮ LIỆU

- Một CSDL được quản lý bởi một hệ quản trị CSDL thường được gọi là một ***hệ cơ sở dữ liệu***.

- Hệ CSDL gồm 4 thành phần:

1. CSDL hợp nhất: có 2 tính chất là tối thiểu hóa dư thừa và được chia sẻ.
2. Người dùng: là những người có nhu cầu truy nhập vào CSDL (người dùng cuối, người viết chương trình ứng dụng, người quản trị CSDL).
3. Phần mềm hệ quản trị CSDL.
4. Phần cứng: gồm các thiết bị nhớ thứ cấp được sử dụng để lưu trữ CSDL.

CÁC HỆ CSDL TRUYỀN THỐNG

- Hệ CSDL thương mại đầu tiên xuất hiện vào những năm 1960. Đó là các hệ thống lưu trữ theo kiểu tệp truyền thống.
- Xét theo 4 đặc tính của hệ quản trị CSDL, các hệ thống tệp:
  - Cung cấp đặc tính (3), tuy nhiên, không hoặc cung cấp rất ít đặc tính (4).
  - Không hỗ trợ trực tiếp đặc tính (2), ví dụ: không hỗ trợ ngôn ngữ truy vấn.
  - Không hỗ trợ đặc tính (1). Chỉ tạo cấu trúc thư mục cho các tệp. Việc hỗ trợ cho các lược đồ rất hạn chế.
- Một số hệ CSDL truyền thống quan trọng hơn, trong đó dữ liệu được chia nhỏ thành các mục, các truy vấn và sửa đổi có thể được thực hiện. Ví dụ: hệ thống bán vé máy bay hoặc hệ thống ngân hàng.

CÁC HỆ CSDL TRUYỀN THỐNG (cont.)

- Phát triển vượt bậc của các hệ CSDL được đề xuất bởi Codd vào năm 1970.
  - CSDL được biểu diễn dưới dạng các bảng (các quan hệ)
  - Cấu trúc dữ liệu phức tạp cho phép đáp ứng nhanh các truy vấn. Người dùng không cần biết đến cấu trúc lưu trữ dữ liệu.
  - Các truy vấn có thể được thể hiện bởi một ngôn ngữ bậc cao, làm tăng hiệu suất cho những người lập trình CSDL.

HỆ THỐNG NGÀY CÀNG NHỎ

- Trước đây:
  - Hệ quản trị CSDL là hệ thống rất lớn, có giá thành cao và chạy trên các máy tính mainframe.
  - Kích cỡ lưu trữ dữ liệu rất lớn nên cần các bộ lưu trữ lớn.
- Ngày nay:
  - Do công nghệ phát triển, một gigabyte có thể được lưu trữ trên một đĩa đơn. Và các hệ quản trị CSDL có thể chạy trên một máy tính cá nhân.

=> Hệ thống ngày càng nhỏ dần theo thời gian do công nghệ điện tử ngày càng phát triển.

=> Hệ quản trị CSDL dựa trên mô hình quan hệ bắt đầu xuất hiện như một công cụ chung cho các ứng dụng máy tính.

DỮ LIỆU NGÀY CÀNG LỚN

- Ngày nay, một gigabyte không còn được coi là dữ liệu có kích cỡ lớn nữa. Các hệ cơ sở dữ liệu lớn phải chứa hàng Terabytes hoặc nhiều hơn.
- Khi bộ nhớ lưu trữ trở nên rẻ và sẵn hơn, con người thường có các lý do mới để lưu trữ nhiều dữ liệu hơn.

  Ví dụ, các cửa hàng bán lẻ thường lưu trữ tới terabytes (1 terabyte = 1000 gigabytes hoặc 10^{12} bytes) thông tin về lịch sử giao dịch mua bán trong một khoảng thời gian rất dài.
- Ngoài dạng văn bản và số, dữ liệu còn có nhiều dạng khác như âm thanh, hình ảnh thường chiếm không gian lưu trữ rất lớn.

  Ví dụ: một giờ của video sẽ chiếm một gigabyte; hay CSDL lưu trữ các hình ảnh vệ tinh sẽ chiếm nhiều petabytes dữ liệu (1 petabyte=1000 Terabytes hay 10^{15} bytes).

=> Xu hướng hiện nay là dữ liệu ngày càng lớn.

DỮ LIỆU NGÀY CÀNG LỚN (cont.)

- Để xử lý được các CSDL lớn đòi hỏi nhiều công nghệ tiên tiến.
  - Các CSDL hầu như không bao giờ cho rằng "dữ liệu" sẽ vừa với bộ nhớ trong. Các hệ thống cũ thường chỉ có các thiết bị lưu trữ thứ cấp dưới dạng các đĩa từ (công nghệ tương tự).
  - CSDL hiện đại được lưu trữ trên một mảng các ổ cứng (các thiết bị lưu trữ thứ cấp).
- Hai xu hướng cho phép các hệ CSDL có thể xử lý được khối lượng dữ liệu lớn một cách nhanh hơn là: ***Lưu trữ mức độ cấp 3*** và ***Tính toán song song***.

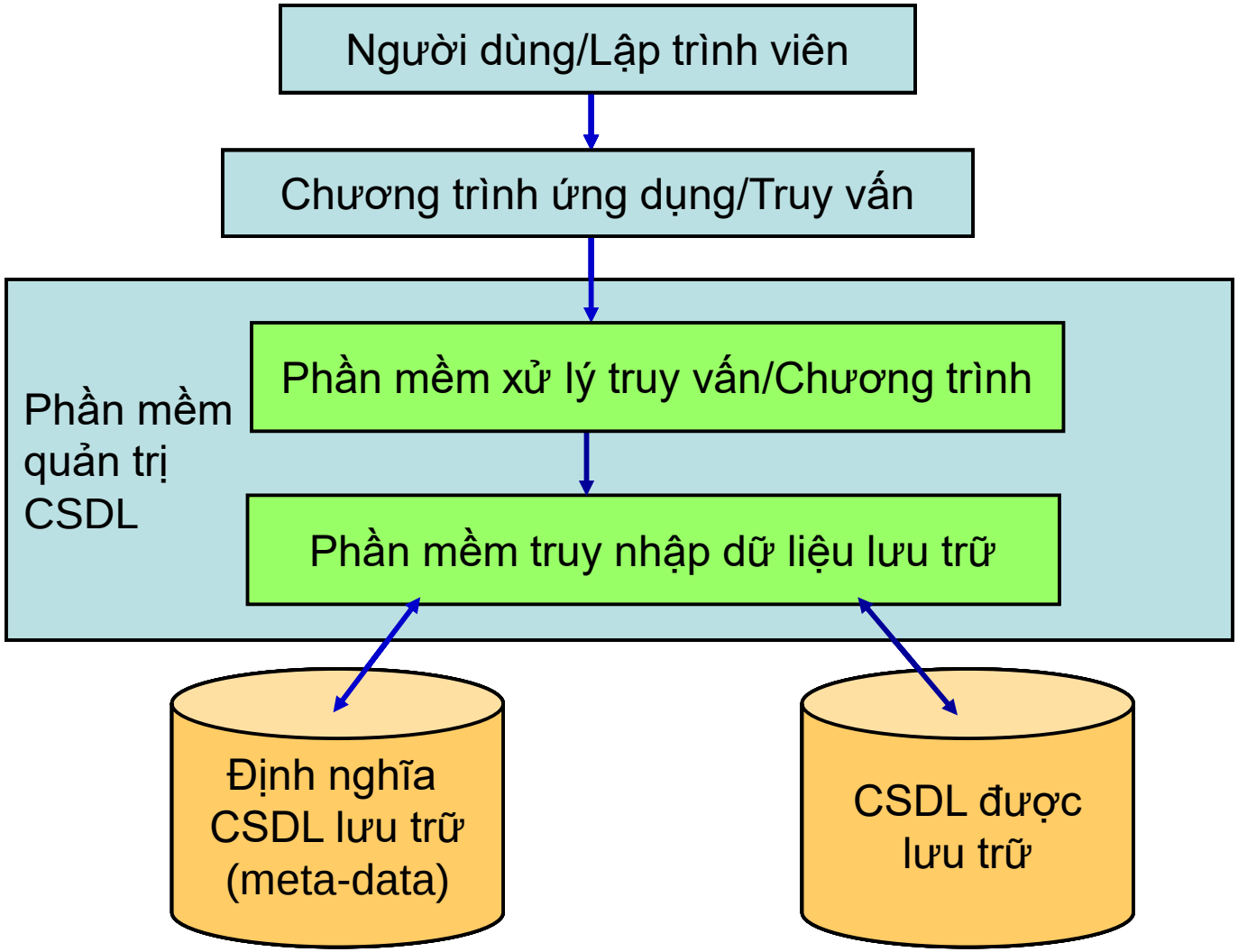
LƯU TRỮ MỨC ĐỘ CẤP 3

- Các CSDL lớn hiện nay đòi hỏi nhiều hơn việc chỉ lưu trữ trên các ổ đĩa (cấp 2). ***Các thiết bị cấp 3*** có xu hướng lưu trữ theo đơn vị terabyte và có thời gian truy nhập dài hơn các ổ đĩa truyền thống.
  - Thời gian truy nhập của một đĩa truyền thống là khoảng 10-20 msec. Trong khi đó của thiết bị cấp 3 là vài giây.
  - Các thiết bị cấp 3 liên quan tới việc chuyển một đối tượng mà trên đó dữ liệu được lưu trữ, tới một thiết bị đọc nào đó.
  - Ví dụ: Đĩa CDs là một phương tiện lưu trữ mức độ cấp 3.

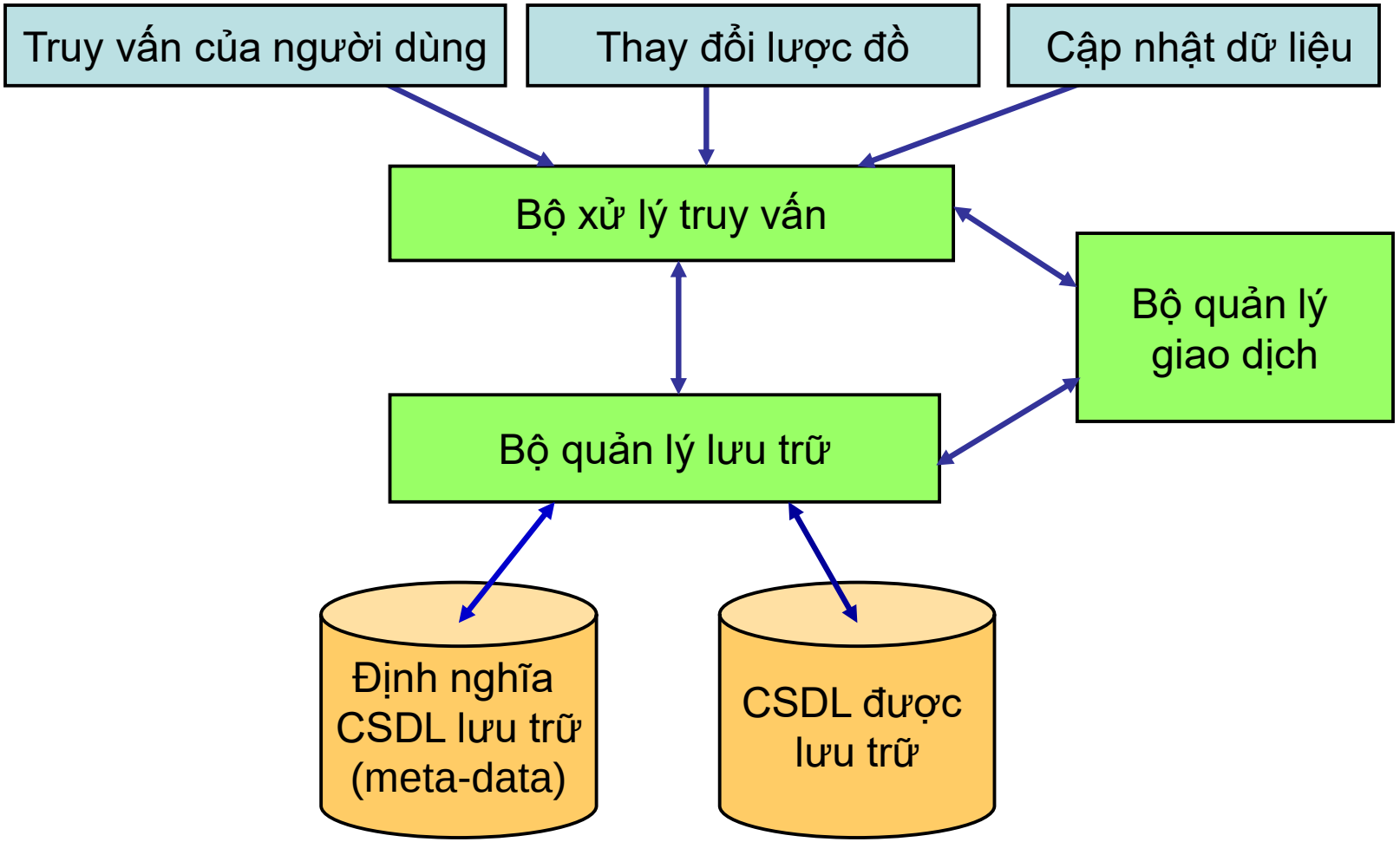
TÍNH TOÁN SONG SONG

- Khả năng lưu trữ khối lượng dữ liệu khổng lồ là rất quan trọng nhưng nó sẽ ít được sử dụng nếu như không thể truy nhập vào khối dữ liệu đó một cách nhanh chóng **=> *Yêu cầu cải thiện tốc độ.***
- Trong CSDL hiện đại, việc cải thiện tốc độ được thực hiện bằng:
  - ***Các cấu trúc chỉ mục***
  - ***Cơ chế song song hóa*** - liên quan tới cả song song hóa bộ vi xử lý và song song hóa bản thân dữ liệu.

TỔNG QUAN CÁC THÀNH PHẦN CỦA MỘT HỆ CSDL



CÁC THÀNH PHẦN CỦA MỘT DBMS (cont.)



Kiến trúc của một hệ quản trị CSDL

CÁC THÀNH PHẦN CỦA MỘT DBMS (cont.)

- ***CSDL lưu trữ và meta-data:***
  - CSDL được lưu trữ tại thiết bị nhớ thứ cấp hoặc cấp 3.
  - Meta-data (siêu dữ liệu) là dữ liệu về dữ liệu: Mô tả các thành phần dữ liệu của CSDL (vị trí tương đối của các trường trong bản ghi, thông tin về lược đồ, thông tin về chỉ mục, ...).
  - Với mỗi CSDL, hệ quản trị CSDL có thể duy trì nhiều chỉ mục khác nhau được thiết kế để cung cấp truy nhập nhanh tới dữ liệu ngẫu nhiên.
  - Trong các CSDL hiện đại, hầu hết các chỉ mục được biểu diễn dưới dạng B-tree (cây tìm kiếm nhị phân). Các B-tree có xu hướng "ngắn và béo" giúp truy nhập nhanh từ gốc đến lá.

CÁC THÀNH PHẦN CỦA MỘT DBMS (cont.)

- ***Bộ quản lý lưu trữ:*** Trong các hệ CSDL đơn giản, bộ quản lý lưu trữ chỉ như là hệ thống tệp trong hệ điều hành. Với các hệ thống lớn hơn, để hiệu quả, hệ quản trị CSDL thường quản lý việc lưu trữ trực tiếp trên ổ đĩa.

  Bộ quản lý lưu trữ có 2 thành phần cơ bản:

  - ***Bộ quản lý tệp:*** Lưu vị trí các tệp trên ổ đĩa và lấy ra được khối hoặc các khối chứa tệp theo yêu cầu từ bộ quản lý vùng đệm.
  - ***Bộ quản lý vùng đệm:*** Quản lý bộ nhớ chính. Lấy các khối dữ liệu từ ổ đĩa, qua bộ quản lý tệp, và chọn một trang trong bộ nhớ chính để lưu trữ. Thuật toán tạo trang sẽ xác định trang sẽ tồn tại bao lâu trong bộ nhớ chính.

CÁC THÀNH PHẦN CỦA MỘT DBMS (cont.)

- ***Bộ xử lý truy vấn:*** Biến đổi một câu truy vấn hoặc một thao tác CSDL, đang được biểu diễn tại một mức rất cao (ví dụ, ngôn ngữ SQL), thành một chuỗi các yêu cầu đối với dữ liệu được lưu trữ trong CSDL.

  Phần phức tạp nhất của bộ xử lý truy vấn là ***tối ưu hóa truy vấn***, nghĩa là chọn ra được chiến lược tốt nhất để thực thi truy vấn.

CÁC THÀNH PHẦN CỦA MỘT DBMS (cont.)

- ***Bộ quản lý giao dịch:*** Giao dịch là một tập các thao tác được xử lý như một đơn vị không chia cắt được. Để đảm bảo được tính chất này, bộ quản lý giao dịch phải đảm bảo 4 tính chất (được gọi là ***thuộc tính ACID***):
  - ***Tính nguyên tố (Atomicity):*** tất cả các thao tác của giao dịch được thực hiện hoặc không thao tác nào được thực hiện.
  - ***Tính nhất quán (Consistency):*** các thao tác phải đảm bảo tính nhất quán của CSDL.
  - ***Tính biệt lập (Isolation):*** các giao dịch đồng thời phải được tách riêng biệt nhau.
  - ***Tính duy trì (Durability):*** những thay đổi tới CSDL bởi một giao dịch sẽ không bị mất đi ngay cả khi hệ thống có lỗi ngay sau khi giao dịch hoàn thành.

CÁC THÀNH PHẦN CỦA MỘT DBMS (cont.)

- ***Ba kiểu thao tác:***
  - ***Truy vấn của người dùng:*** là các thao tác hỏi đáp về dữ liệu được lưu trữ trong CSDL. Chúng được sinh ra theo 2 cách: (1) Thông qua giao diện truy vấn chung, (2) Thông qua giao diện chương trình ứng dụng.
  - ***Cập nhật dữ liệu:*** là các thao tác thay đổi dữ liệu, như thêm, sửa, xóa dữ liệu trong CSDL. Chúng cũng được sinh ra theo 2 cách (1) và (2) như trên.
  - ***Thay đổi lược đồ:*** là các lệnh được sinh ra bởi người dùng được cấp phép, thường là người quản trị CSDL.

DỮ LIỆU VÀ THÔNG TIN

- Xét ví dụ về dữ liệu gồm các con số như sau:

Data:

0	11,500
5	12,300
10	12,800
15	10,455
20	12,200
25	13,900
30	14,220

- Thể hiện dưới dạng “thô” như trên, dữ liệu có rất ít ý nghĩa. Nó đơn giản là một cặp danh sách các số nguyên. Không có ngữ cảnh làm nền cho dữ liệu.

DỮ LIỆU VÀ THÔNG TIN (cont.)

- ***Chuyển dữ liệu thành một dạng có ý nghĩa hơn:***

  Quá trình xử lý cơ bản là đặt dữ liệu vào trong ngữ cảnh *(thường được thực hiện bằng cách thêm dữ liệu vào. Mặc dù những dữ liệu thêm vào thực sự là siêu dữ liệu).*

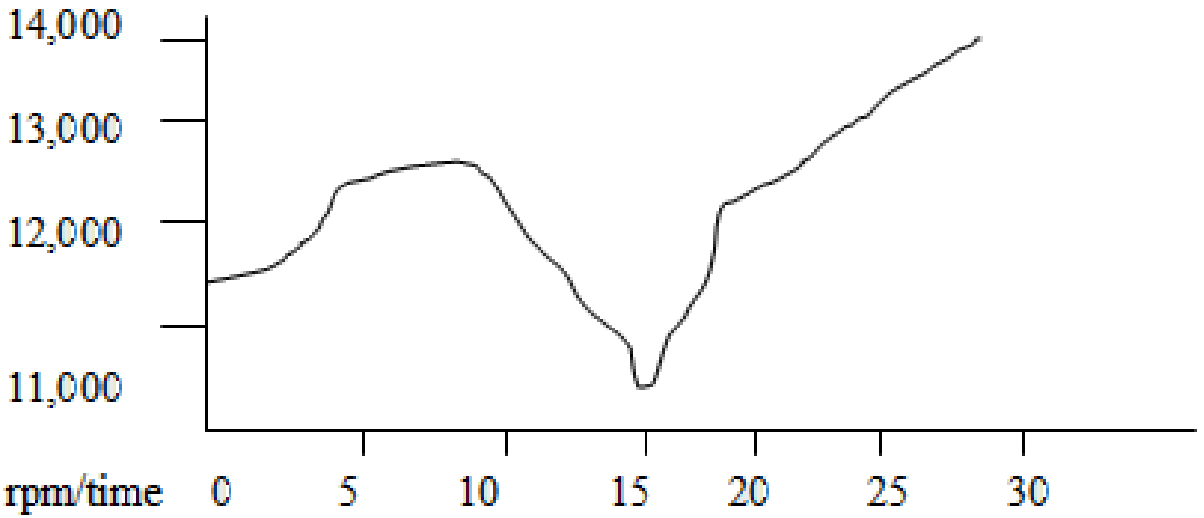
- ***Dữ liệu bắt đầu có ý nghĩa hơn như sau:***

Thông tin: Dữ liệu Engine RPM: Roebling Road 10/4/2003 – Yamaha Heavy

Lap 12:	**time**	**rpm**
	0	**11,500**
	5	**12,300**
	10	**12,800**
	15	**10,455**
	20	**12,200**
	25	**13,900**
	30	**14,220**

DỮ LIỆU VÀ THÔNG TIN (cont.)

- ***Cùng với dữ liệu trên, xem xét quá trình xử lý chuyển dữ liệu thành dạng đồ thị:***



Đồ thị: Một phần Lap 12 - Roebling Road 10/4/2003 – Yamaha Heavy

DỮ LIỆU DẪN XUẤT VÀ DỮ LIỆU VẬT LÝ

- ***Dữ liệu vật lý:*** là những dữ liệu có thực, được nhập vào trong CSDL.
- ***Dữ liệu dẫn xuất:*** Là những dữ liệu được tính toán từ những dữ liệu nằm trong CSDL.
- Phụ thuộc vào mức độ phức tạp của chương trình ứng dụng và hệ quản trị CSDL, khối dữ liệu dẫn xuất có thể lớn hơn rất nhiều khối dữ liệu vật lý.
- Cân nhắc khi nào dữ liệu dẫn xuất trở thành dữ liệu vật lý?

  Ví dụ: CSDL lưu thông tin về Sinh viên, trong đó lưu điểm thi của SV. Giá trị trung bình điểm thi của SV có cần lưu trong CSDL hay sẽ được tính toán khi cần?