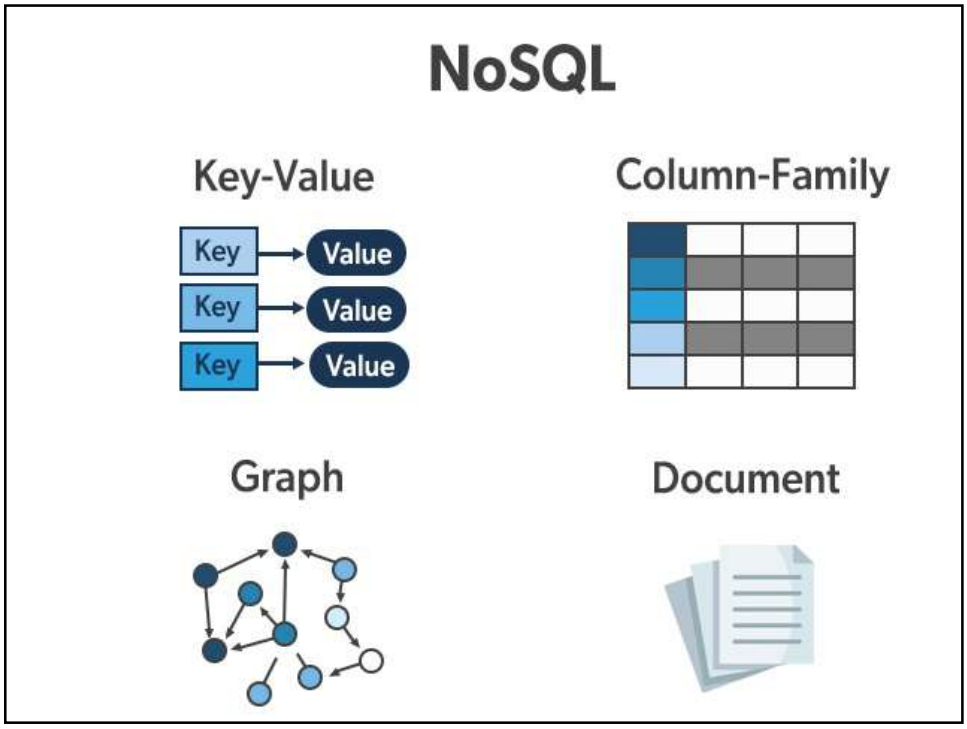


TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHỆ CẦN THƠ
KHOA: CÔNG NGHỆ THÔNG TIN

GIỚI THIỆU VỀ MÔN HỌC
HỆ QUẢN TRỊ CSDL NO SQL

Giảng viên: Nguyễn Bá Duy
Tel: 0983.877750
Email: nbduy@ctuet.edu.vn

KẾ HOẠCH GIẢNG DẠY

- Số tín chỉ: 2TC
- Hình thức:
  - Trình bày.
  - Bài tập.

ĐÁNH GIÁ GIỮA KỲ (40%)

Bài tập nhóm

THI CUỐI KỲ (60%)

- Trắc nghiệm
- 40 câu
- 45 phút
- 2 đề
- Được sử dụng tài liệu

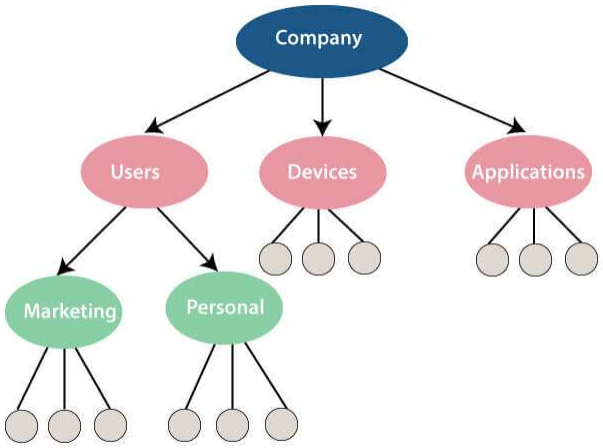
Giới thiệu môn học

- Chương 1: Sơ lược về DBMS.
- Chương 2: Giới thiệu về NoSQL.
- Chương 3: Các loại CSDL NoSQL.
- Chương 4: Giới thiệu MongoDB.
- Chương 5: Ứng dụng CSDL NoSQL.

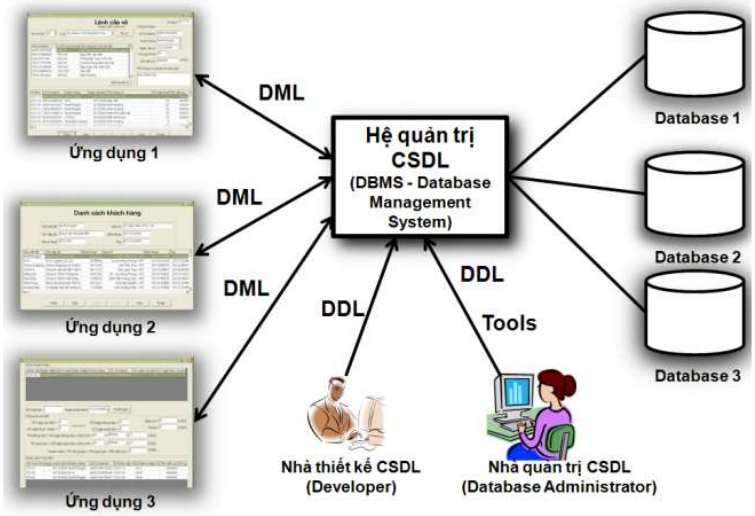
CÁCH MẠNG CÔNG NGHIỆP

CÁCH MẠNG CÔNG NGHIỆP 4.0

CƠ SỞ DỮ LIỆU



HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU



QUẢN LÝ DỮ LIỆU

- Quan niệm truyền thống:

Quản lý dữ liệu đồng nghĩa với hệ thống quản lý cơ sở dữ liệu quan hệ (Relational Database Management System - RDBMS)

ORACLE DB Management

MySQL

Microsoft SQL Server

Ví dụ: Nếu chúng ta muốn chuyển $100 từ tài khoản tiết kiệm sang tài khoản séc của mình, cần thực hiện hai bước:

- Trích ra $100 từ tài khoản tiết kiệm;
- Thêm $100 vào tài khoản séc.

=> Câu hỏi đặt ra: Trường hợp thực hiện xong bước 1, có sự cố xãy ra và không thực hiện bước 2 thì chúng ta có mất $100 không?

=> Nếu chúng ta sử dụng hệ quản trị cơ sở dữ liệu quan hệ, điều này có thể không xảy ra nếu chúng ta sử dụng Transaction <= CSDL quan hệ có thể nhóm một tập hợp các hoạt động, chẳng hạn như khấu trừ từ khoản tiết kiệm và thêm vào tài khoản séc, như một hoạt động duy nhất. Nếu một trong những hoạt động không thực hiện được, RDBMS sẽ không thực hiện hành động đối với dữ liệu truyền vào.

=> Không mất tiền.

DỮ LIỆU TRONG THỜI ĐẠI MỚI

CƠ SỞ DỮ LIỆU NOSQL

CÁC HỆ THỐNG QUẢN LÝ DỮ LIỆU SƠ KHAI

Trước năm 1970:

- Hệ thống quản lý dữ liệu tập tin
- Hệ thống quản lý dữ liệu phân cấp
- Hệ thống quản lý dữ liệu mạng

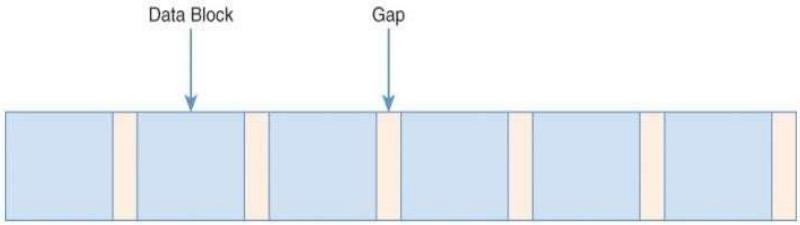
CÁC HỆ THỐNG QUẢN LÝ DỮ LIỆU SƠ KHAI

1. Hệ thống quản lý dữ liệu tập tin

Các tập tin (*file*) là một tập hợp dữ liệu có tổ chức được lưu trữ trên một phương tiện lưu trữ bền vững như đĩa hoặc băng từ.

Các tập tin dữ liệu phải đáp ứng các ràng buộc của hệ thống lưu trữ vật lý.

Băng từ là một vật liệu nhựa dài, mỏng có từ tính và là một phương tiện phổ biến để ghi âm thanh từ những năm 1950 đến những năm 1970. Nó cũng được điều chỉnh để lưu trữ dữ liệu kỹ thuật số. Một băng từ được chia thành một loạt các khối với các khoảng trống (*gap*) giữa chúng.



Băng từ lưu trữ dữ liệu trong các khối tuần tự

Trong đó:

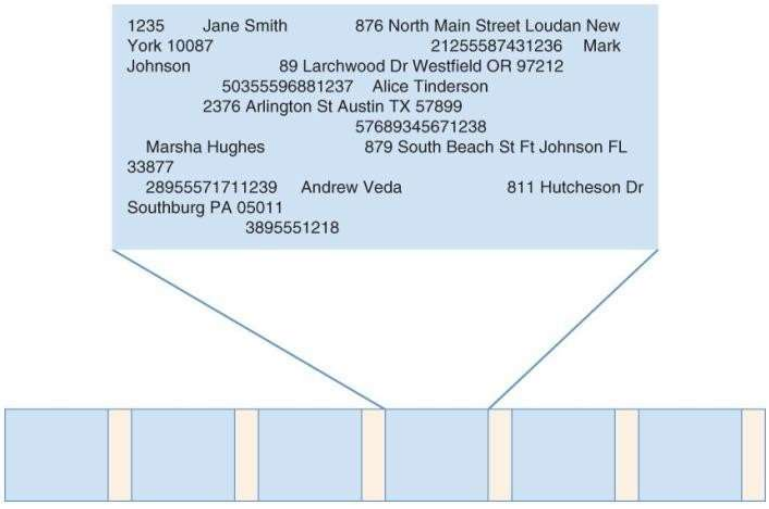
- Data Block: là vị trí chứa dữ liệu.
- Gap: Đánh dấu kết thúc/bắt đầu 1 data block.

Dữ liệu được ghi vào các khối bằng các đầu ghi trong ổ băng. Dữ liệu cũng được đọc bằng cách di chuyển băng qua đầu đọc/ghi.

Ví dụ: Lưu trữ dữ liệu khách hàng:

- Customer ID - 10 ký tự
- Customer name - 40 ký tự
- Customer address - 100 ký tự
- Customer phone number - 10 ký tự

Để lưu trữ thông tin của mỗi khách hàng, cần có 160 ký tự. Nếu một khối trên băng dài 800 ký tự, chúng ta có thể lưu trữ năm bản ghi khách hàng trong mỗi khối.



Khối là một phần dữ liệu được đọc trong một thao tác đọc duy nhất

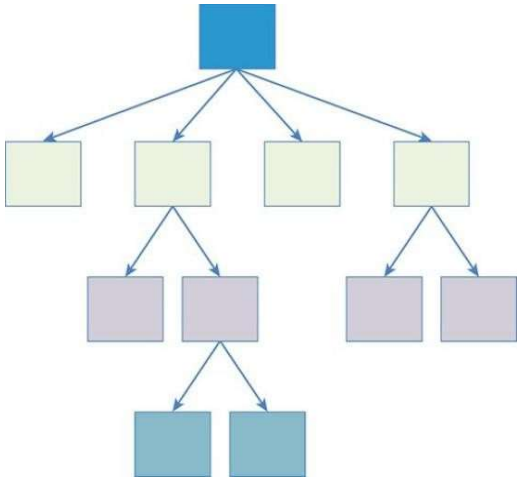
Hạn chế:

- Các chương trình sử dụng tập tin quyết định phần lớn việc tổ chức dữ liệu. Ví dụ: Chúng ta muốn tổ chức tập tin theo hồ sơ khách hàng được sắp xếp theo ID khách hàng. Điều này giúp cho việc bổ sung khách hàng mới hiệu quả hơn. Khi mỗi khách hàng mới được tạo, khách hàng có thể được thêm vào cuối băng từ.
- Khó tìm kiếm dữ liệu.

- Có thể dẫn đến dữ liệu bị trùng lặp. Do cấu trúc lưu trữ không nhất quán, dẫn đến có thể lưu trữ trùng lặp dữ liệu (do không có cơ chế kiểm tra trước khi thêm dữ liệu mới vào).
- Khó chia sẻ tập tin đối với trường hợp thông tin cần được giữ bí mật với một số người dùng. Ví dụ: Một tập tin nhân viên chứa tên, địa chỉ, số điện thoại, ID nhân viên và chức danh của tất cả nhân viên sẽ hữu ích cho một số bộ phận khác nhau của tổ chức. Tuy nhiên, nếu tập tin cũng chứa thông tin về tiền lương, thì dữ liệu đó chỉ nên được truy cập bởi những người có trách nhiệm công việc yêu cầu nó. *Phân quyền?*

2. Mô hình quản lý dữ liệu phân cấp

Một trong những hạn chế của hệ thống quản lý dữ liệu dựa trên tập tin là chúng có thể không hiệu quả để tìm kiếm. Mô hình dữ liệu phân cấp giải quyết vấn đề này bằng cách tổ chức dữ liệu trong một hệ thống phân cấp của các mối quan hệ cha-con.



Mô hình phân cấp được tổ chức thành một tập hợp các quan hệ cha-con

Nút (Node): Đại diện cho một thực thể dữ liệu trong cây, được kết nối với các nút khác bằng các liên kết (links) biểu thị mối quan hệ giữa chúng.

- Nút gốc (root node): Là nút trên cùng của cây, không có nút
- Nút cha (parent node): Là nút có nút con kết nối với nó.
- Nút con (child node): Là nút kết nối với nút cha của nó.
- Nút lá (leaf node): Là nút không có nút con.

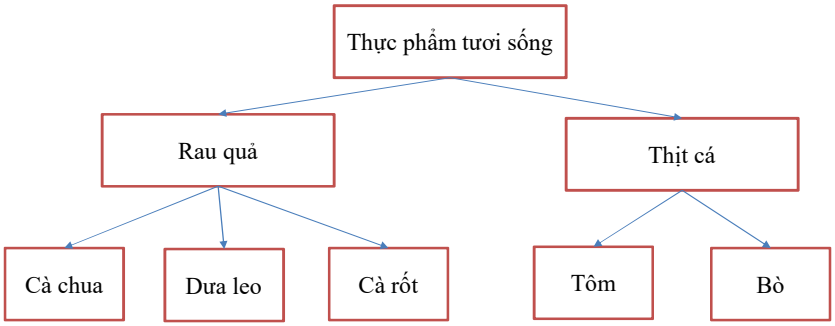
Mức (level): Đại diện cho độ sâu của một nút trong cây, tính bằng cách đếm số lần đi từ nút gốc đến nút đó.

Các liên kết (links): Biểu thị mối quan hệ giữa các nút trong cây.

Tập hợp các nút con (children set): Là tập hợp các nút trực tiếp kết nối với nút cha.

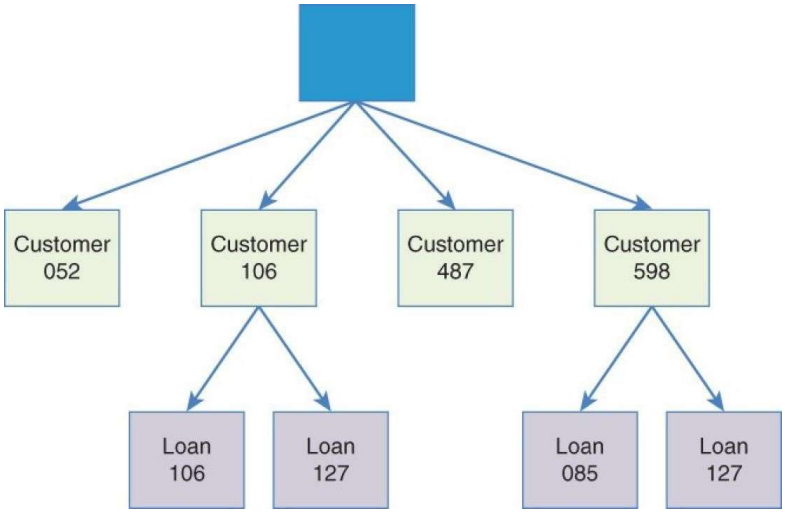
Đường đi (path): Là tập hợp các nút trên một đường từ nút cha đến nút con.

Ví dụ 1: Mô hình phân cấp thực phẩm



Ví dụ 2:

Xem xét dữ liệu mà bộ phận cho vay của ngân hàng có thể theo dõi. Ngân hàng có nhiều khách hàng và mỗi khách hàng có một hoặc nhiều khoản vay. Đối với mỗi khách hàng, bộ phận cho vay sẽ muốn theo dõi tên, địa chỉ và số điện thoại của khách hàng. Đối với mỗi khoản vay, bộ phận cho vay cần theo dõi số tiền vay, lãi suất, ngày khoản vay được thực hiện và ngày khoản vay đến hạn. Khách hàng có thể có nhiều khoản vay cùng một lúc và một khoản vay có thể có nhiều khách hàng liên kết với nó.



Mô hình dữ liệu phân cấp
cho cơ sở dữ liệu quản lý khoản vay

Một lợi thế của mô hình phân cấp so với các tập tin là tìm kiếm hiệu quả hơn. Thay vì phải quét tất cả dữ liệu trên băng từ để tìm kiếm một khối dữ liệu, một chương trình sử dụng mô hình phân cấp có thể chỉ quét hồ sơ khách hàng để tìm kiếm hồ sơ khoản vay của một khách hàng cụ thể. Khi hồ sơ khách hàng được tìm thấy, chương trình có thể tìm kiếm thông qua các khoản vay của khách hàng để tìm khoản vay có lãi suất cụ thể.

Hạn chế:

- Khó biểu diễn quan hệ nhiều cha – 1 con. Ví dụ: Hai khách hàng với một khoản vay, chẳng hạn như hai đối tác kinh doanh cùng vay một khoản vay kinh doanh ngắn hạn.
- Dữ liệu có thể sai lệch: không cẩn thận có thể dẫn đến dữ liệu không nhất quán. Ví dụ: thêm 1 khoản vay giống nhau cho 2 khách hàng, hệ thống tạo 2 khoản vay khác nhau giữa 2 khách hàng.

- Có khả năng xảy ra lỗi khi tổng hợp dữ liệu. Ví dụ: để tìm tổng giá trị của tất cả các khoản vay chưa thanh toán.

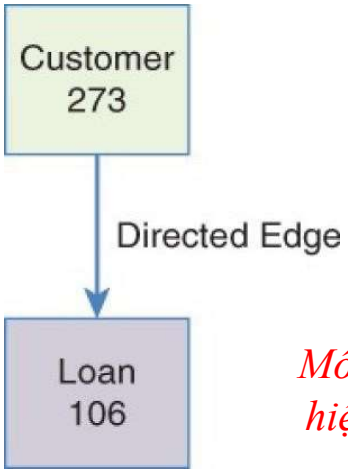
=> Lập trình viên không chỉ đọc tất cả hồ sơ khoản vay và cộng tất cả các khoản vay lại với nhau do 1 khoản vay có thể dành cho 2 khách hàng (cùng vay) có thể dẫn đến tổng số tiền vay lớn hơn số tiền thực tế, do đó lập trình viên phải thực hiện các bước đếm mỗi khoản vay một lần duy nhất.

3. Mô hình quản lý dữ liệu mạng

Mô hình dữ liệu mạng giống như mô hình dữ liệu phân cấp ở chỗ sử dụng các liên kết giữa các bản ghi; tuy nhiên, không giống như các mô hình dữ liệu phân cấp, mô hình này không bị hạn chế chỉ có một bản ghi cha. Ngoài ra, không giống như hệ thống quản lý dữ liệu tập tin và hệ thống quản lý dữ liệu phân cấp, các mô hình dữ liệu mạng có hai thành phần chính yếu: một lược đồ và cơ sở dữ liệu.

Một mạng được tạo thành từ các bản ghi dữ liệu được liên kết với nhau. Các bản ghi dữ liệu được gọi là các nút (*node*) và các liên kết được gọi là các cạnh (*edge*). Tập hợp các nút và các cạnh được gọi là một đồ thị (*graph*).

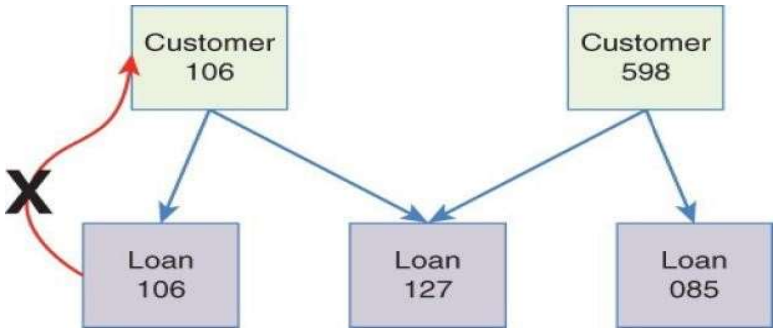
Cho phép biểu diễn các mối quan hệ cha - con. Quan hệ cha - con còn được gọi là quan hệ một - nhiều hoặc nhiều nhiều.



Mối quan hệ cha-con được thể hiện bằng một cạnh có hướng

Có 2 ràng buộc: Cạnh phải có hướng đến và không được có chu trình.

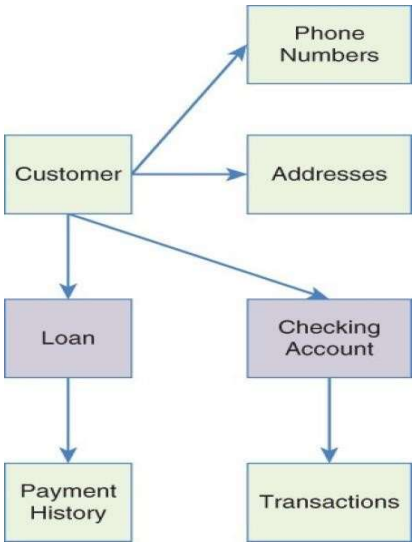
Đồ thị có cạnh có hướng và không có chu trình được gọi là đồ thị xoay chiều có hướng (*directed acyclic graph*).



Chu trình trong mô hình quản lý dữ liệu mạng

Các loại nút có thể liên kết với các nút khác được xác định trong một cấu trúc được gọi là lược đồ (*schema*).

Một lược đồ mạng cho thấy các thực thể có thể liên kết với các thực thể khác.



Phần khác của hệ thống quản lý dữ liệu mạng là chính cơ sở dữ liệu. Đây là nơi dữ liệu thực tế được lưu trữ theo cấu trúc của lược đồ. Một trong những tiến bộ của cơ sở dữ liệu mạng so với các cách tiếp cận trước đó là nó đã được tiêu chuẩn hóa vào năm 1969 bởi Hội nghị về Ngôn ngữ Hệ thống Dữ liệu (*Conference on Data Systems Languages - CODASYL*) Consortium. Tiêu chuẩn này đã trở thành cơ sở cho hầu hết các triển khai cơ sở dữ liệu mạng.

Hạn chế:

- Khó thiết kế và bảo trì. Tùy thuộc vào cách các nút được liên kết, một chương trình có thể cần phải đi qua một số lượng lớn các liên kết để đến nút chứa dữ liệu cần thiết. Khi các mô hình dữ liệu trở nên phức tạp hơn, số lượng liên kết và độ dài tăng lên đáng kể.

- Sau khi cơ sở dữ liệu mạng được triển khai, người thiết kế cơ sở dữ liệu thêm 1 nút, thì các chương trình truy cập cơ sở dữ liệu mạng sẽ phải được cập nhật (thêm nút mới). Việc thêm các nút vào lược đồ và cơ sở dữ liệu sẽ thay đổi các đường đi mà chương trình phải duyệt qua để đến các nút cụ thể.

HẠN CHẾ CỦA CÁC HT QLDL SƠ KHAI

- Dữ liệu trùng lặp
- Khó thực hiện bảo mật
- Tìm kiếm không hiệu quả
- Khó bảo trì mã chương trình sử dụng cơ sở dữ liệu.

=> Tính độc lập về cấu trúc của việc tổ chức logic và vật lý của cơ sở dữ liệu là một bước tiến lớn trong quản lý dữ liệu được cung cấp bởi các hệ quản trị cơ sở dữ liệu quan hệ.