

TRƯỜNG ĐẠI HỌC THĂNG LONG
KHOA TOÁN TIN



CƠ SỞ PHÂN TÁN

**ĐỀ TÀI: XÂY DỰNG CƠ SỞ DỮ LIỆU PHÂN TÁN CHO BÀI
TOÁN QUẢN LÝ SINH VIÊN**

GIÁO VIÊN HƯỚNG DẪN

Đậu Hải Phong

SINH VIÊN THỰC HIỆN

Tên nhóm: Nhóm 1

Đào Tuấn Trung – A28563

Đặng Anh Tú - A29378

Ngành: Khoa học máy tính

HÀ NỘI – 2019

LỜI GIỚI THIỆU

Tài liệu này là tài liệu báo cáo bài tập lớn môn Cơ sở dữ liệu phân tán Học kỳ II nhóm 2 năm học 2019-2020. Tài liệu mô tả lại toàn bộ quy trình xây dựng cơ sở dữ liệu phân tán cho bài toán Quản lý sinh viên. Tài liệu bao gồm 3 chương chính:

- Chương 1: Tổng quan về cơ sở dữ liệu phân tán: Chương này giúp người người xem có cách nhìn tổng quát về cơ sở dữ liệu phân tán
- Chương 2: Xây dựng cơ sở dữ liệu phân tán cho bài toán Quản lý sinh viên: Chương này mô tả về bài toán Quản lý sinh viên và mô hình dữ liệu quan hệ của bài toán
- Chương 3: Thiết kế và Cài đặt cơ sở dữ liệu phân tán: Chương này trình bày chi tiết về thiết kế và cài đặt cơ sở dữ liệu phân tán trong bài toán Quản lý sinh viên

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN VỀ CƠ SỞ DỮ LIỆU PHÂN TÁN 4

- 1.1. Cơ sở dữ liệu phân tán là gì? 4
 - 1.1.1. Khái niệm 4
 - 1.1.2. Đặc điểm 4
 - 1.1.3. Ưu điểm và hạn chế..... 5
- 1.2. Kiến trúc cơ sở dữ liệu phân tán 6
- 1.3. Khả năng của cơ sở dữ liệu phân tán 8
- 1.4. Các vấn đề cần quan tâm trong cơ sở dữ liệu phân tán 8

Chương 2: XÂY DỰNG CƠ SỞ DỮ LIỆU PHÂN TÁN CHO BÀI TOÁN QUẢN LÝ SINH VIÊN 10

- 2.1. Mô tả bài toán 10
- 2.2. Mô hình dữ liệu quan hệ 11

Chương 3: THIẾT KẾ VÀ CÀI ĐẶT CƠ SỞ DỮ LIỆU PHÂN TÁN..... 12

- 3.1. Thiết kế cơ sở dữ liệu phân tán 12
 - 3.1.1. Phân tích phân mảnh dọc (thuật toán BEA):..... 12
 - 3.1.2. Phân tích phân mảnh ngang (hội vị từ) 14
- 3.2. Xây dựng dữ liệu từ điển dữ liệu toàn diện 15
 - 3.2.1. Global Conceptual Schema: 15
 - 3.2.2. Data Directory 15
 - 3.2.3. Fragmentation Directory: 15
 - 3.2.4. Kiểm soát dữ liệu 16
- 3.3. Xây dựng các hàm tính năng phụ trợ:..... 17
 - 3.3.1. Tính điểm trung bình của một sinh viên 17
 - 3.3.2. Lấy bảng điểm của một sinh viên 17
 - 3.3.3. Lấy tổng số tín chỉ của một sinh viên 17

CHƯƠNG 1: TỔNG QUAN VỀ CƠ SỞ DỮ LIỆU PHÂN TÁN

1.1. Cơ sở dữ liệu phân tán là gì?

1.1.1. Khái niệm

- Cơ sở dữ liệu phân tán một tập hợp gồm nhiều CSDL được kết nối với nhau, được phân tán về mặt vật lý ở các vị trí và giao tiếp qua mạng
- Người dùng truy cập vào CSDL phân tán thông qua các chương trình ứng dụng. Các chương trình ứng dụng được chia làm 2 loại:
 - Chương trình không yêu cầu dữ liệu từ nơi khác
 - Chương trình có yêu cầu dữ liệu từ nơi khác
- Hệ quản trị CSDL phân tán là một hệ thống phần mềm cho phép quản trị CSDL phân tán và làm cho người sử dụng không nhận thấy sự phân tán về lưu trữ dữ liệu
- Hệ CSDL phân tán có thể chia làm 2 loại chính:
 - Hệ CSDL phân tán thuần nhất: các nút trên mạng đều dùng cùng một hệ QTCSDL
 - Hệ CSDL phân tán hỗn hợp: các nút trên mạng có thể dùng các hệ QTCSDL khác nhau

1.1.2. Đặc điểm

Cơ sở dữ liệu phân tán không đơn giản là sự phân bố của các CSDL bởi vì CSDL phân tán có nhiều đặc điểm khác biệt so với CSDL tập trung truyền thống. Phần này so sánh CSDL phân tán với các CSDL tập trung ở một số đặc điểm: điều khiển tập trung, sự độc lập dữ liệu, sự giảm thừa dữ liệu, cấu trúc vật lý phức tạp để truy xuất hiệu quả.

1.1.2.1. Điều khiển tập trung

- Trong CSDL phân tán việc điều khiển dữ liệu tập trung ít được tập trung đến hơn, nó còn tùy thuộc vào cấu trúc của CSDL.
- Ở CSDL phân tán việc điều khiển sẽ dựa vào việc điều khiển các mảnh dữ liệu thông qua việc quản lý CSDL cục bộ (do Người quản trị CSDL cục bộ thực hiện) và quản lý CSDL toàn cục (do Người quản trị CSDL toàn cục thực hiện), tại mỗi địa phương mà csdl đang ở đó sẽ mang Tính tự trị vị trí / địa phương tức tại đó họ có thể tự quản lý cơ sở dữ liệu của mình mà không phụ thuộc.

1.1.2.2. Độc lập dữ liệu

Trong CSDL phân tán có một khái niệm mới về Độc lập dữ liệu đó là Tính trong suốt dữ liệu (Data transparency)

- Trong suốt phân mảnh: Không nhìn thấy các mảnh; nhìn thấy các quan hệ toàn cục; lược đồ toàn cục
- Trong suốt vị trí: Không nhìn thấy quan hệ cục bộ, thấy các mảnh; lược đồ phân mảnh
- Trong suốt nhân bản: Nhìn thấy các mảnh; không thấy nhân bản các mảnh
- Trong suốt ánh xạ cục bộ: Nhìn thấy quan hệ cục bộ; không nhìn thấy CSDL vật lý
- Trong suốt phân tán (Distribution transparency): Gồm 4 trong suốt trên

1.1.2.3. Giảm dư thừa dữ liệu

- Giảm dư thừa dữ liệu trong csdl phân tán là phức tạp hơn so với csdl cục bộ vì tính chất phân tán của cơ sở dữ liệu sẽ dẫn đến:
- Hoạt động của các chương trình ứng dụng có thể bị tăng lên khi dữ liệu được sao lại tất cả các vị trí, nơi trình ứng dụng cần nó.
- Khi cập nhật thông tin hay một mẫu tính thì phải cập nhật trên tất cả các site để đảm bảo tính nhất quán của dữ liệu.

1.1.2.4. Độ tin cậy qua các giao dịch phân tán

- Trong CSDL phân tán với một mức độ tự trị rất cao của các địa phương, người chủ dữ liệu địa phương cảm giác được bảo vệ tốt hơn vì họ có thể tự chủ thực hiện bảo vệ thay vì phụ thuộc vào người quản trị CSDL trung tâm.
- Vấn đề bảo mật là bản chất trong hệ phân tán nói chung, vì các mạng truyền thông diện rộng cho phép nhiều người cập nhật và khai thác dữ liệu nên cần được bảo vệ.

1.1.2.5. Dễ dàng mở rộng hệ thống

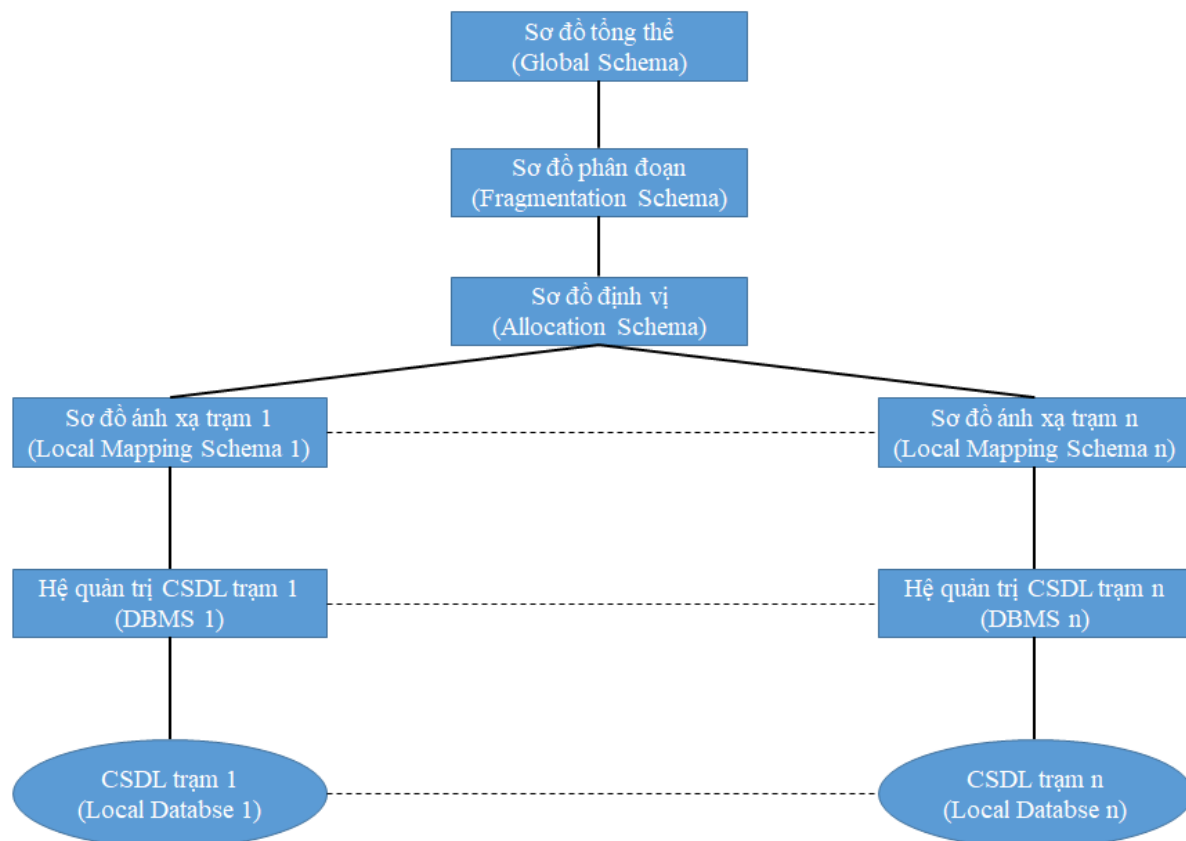
Trong môi trường phân tán, dễ dàng tăng kích thước dữ liệu và hiếm khi cần sửa đổi trong các hệ thống lớn. Việc mở rộng thường có thể thực hiện được bằng các tăng khả năng lưu trữ và xử lý của mạng. Rõ ràng là không thể có được sự gia tăng khả năng một cách tuyến tính vì điều này phụ thuộc vào chi phí phân tán. Tuy nhiên vẫn có thể có những cải tiến có ý nghĩa. Khả năng mở rộng hệ thống dễ dàng mang tính kinh tế, chi phí giảm

1.1.3. Ưu điểm và hạn chế

- Ưu điểm: Sự phân tán dữ liệu và các ứng dụng có một số ưu điểm so với các hệ CSDL tập trung:
 - Cấu trúc phân tán dữ liệu thích hợp cho bản chất phân tán của nhiều người dùng
 - Dữ liệu được chia sẻ trên mạng nhưng vẫn cho phép quản trị dữ liệu tại mỗi trạm

- Dữ liệu có tính sẵn sàng cao
- Dữ liệu có tính tin cậy cao vì khi một nút gặp sự cố, có thể khôi phục dữ liệu tại đây do bản sao của nó có thể được lưu trữ tại một nút khác nữa.
- Hiệu năng của hệ thống được nâng cao hơn
- Cho phép mở rộng các tổ chức một cách linh hoạt. Có thể thêm nút mới vào mạng máy tính mà không ảnh hưởng đến hoạt động của các nút sẵn có
- So với các hệ CSDL tập trung, hệ CSDL phân tán có một số hạn chế như sau:
 - Hệ thống phức tạp hơn vì phải làm ẩn đi sự phân tán dữ liệu đối với người dùng
 - Chi phí cao hơn
 - Đảm bảo an ninh khó khăn hơn
 - Đảm bảo tính nhất quán dữ liệu khó hơn
 - Việc thiết kế CSDL phân tán phức tạp hơn

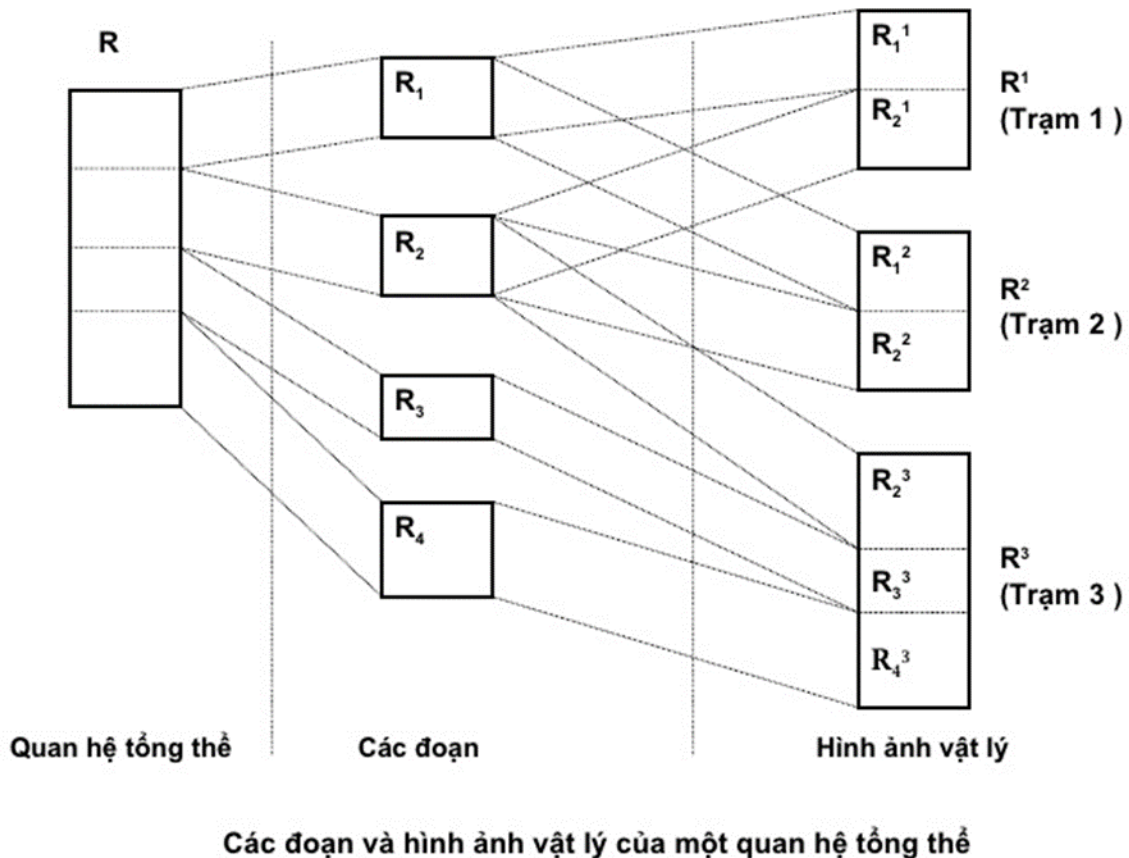
1.2. Kiến trúc cơ sở dữ liệu phân tán



Kiến trúc tham dùng cho CSDL phân tán

- **Sơ đồ tổng thể (Global Schema):**

- Xác định tất cả dữ liệu sẽ được lưu trữ trong CSDL phân tán cũng như các dữ liệu không được phân tán ở các trạm trong hệ thống
- Sơ đồ tổng thể được định nghĩa theo cách như trong CSDL tập trung
- Trong mô hình quan hệ, sơ đồ tổng thể bao gồm định nghĩa của tập các quan hệ tổng thể (Globle relation)
- **Sơ đồ phân đoạn (Fragmentation Schema):**
 - Mỗi quan hệ tổng thể có thể chia thành một vài phần không giao nhau gọi là phân đoạn
 - Có nhiều cách khác nhau để thực hiện việc phân chia này
 - Sơ đồ phân đoạn mô tả các ánh xạ giữa các quan hệ tổng thể và các đoạn được định nghĩa trong sơ đồ phân đoạn
 - Các đoạn được mô tả bằng tên của quan hệ tổng thể cùng với chỉ mục đoạn
- **Sơ đồ định vị (Allocation Schema):**
 - Các đoạn là các phần logic của một quan hệ tổng thể được định vị vật lý trên một hay nhiều trạm
 - Sơ đồ định vị xác định đoạn dữ liệu nào được định vị tại trạm nào trên mạng
 - Tất cả các đoạn được liên kết với cùng một quan hệ tổng thể R và được định vị tại cùng một trạm j cấu thành ảnh vật lý quan hệ tổng thể R tại trạm j
 - Do đó ta có thể ánh xạ 1-1 giữa một ảnh vật lý và một cặp (quan hệ tổng thể, trạm)
 - Các ảnh vật lý có thể chỉ ra bằng tên của một quan hệ tổng thể và một chỉ mục trạm
- **Sơ đồ ánh xạ trạm (Local Mapping Schema):**
 - Thực hiện ánh xạ các ảnh vật lý lên các đối tượng được thực hiện bởi hệ quản trị CSDL trạm
 - Tất cả các đoạn của một quan hệ tổng thể trên cùng một trạm tạo ra một ảnh vật lý



1.3. Khả năng của cơ sở dữ liệu phân tán

- Thực tế tổ chức và kinh tế - Không tập trung tại một nơi mà phân bố trên nhiều vùng địa lý khác nhau
- Các CSDL hiện tại cần kết nối với nhau: Nhiều CSDL đã tồn tại trong 1 công ty và cần được thực hiện nhiều ứng dụng toàn cục
- Sự lớn mạnh tăng: Có thêm các đơn vị tổ chức độc lập
- Giảm chi phí truyền thông: Nhiều ứng dụng cục bộ làm giảm chi phí truyền thông so với CSDL tập trung
- Hiệu suất: Cơ chế song song hóa; Phân mảnh theo ứng dụng, cực đại hóa tính cục bộ ứng dụng
- Độ tin cậy và tính sẵn sàng: Dư thừa dữ liệu; Cần đảm bảo tính tin cậy của dữ liệu

1.4. Các vấn đề cần quan tâm trong cơ sở dữ liệu phân tán

- Scalability(tính dãn trải có thể mở rộng và phát triển rộng khắp)
- Geographic distribution of data(có thể phân bố dữ liệu trên nhiều vùng địa lý,triển khai được trên các vùng cách xa nhau)

- Data “clusters”: (có gom nhóm dữ liệu hay không(mặt vật lý(cluster là dạng index trên ý thức vật lý)))
- Performance: (khả năng thể hiện của cấu trúc)
- Cost: (chi phí)
- CPU: thời gian thực hiện truy suất
- I/O: xuất nhập dữ liệu(memory,disk,..)
- Truyền tải dữ liệu trên hệ thống mạng(đây là đặt điểm quan trọng của việc chọn một cấu trúc cho việc sử dụng csdl phân tán)
- Reliability: (tính bền vững và độ tinh cậy của cấu trúc)

Chương 2: XÂY DỰNG CƠ SỞ DỮ LIỆU PHÂN TÁN CHO BÀI TOÁN QUẢN LÝ SINH VIÊN

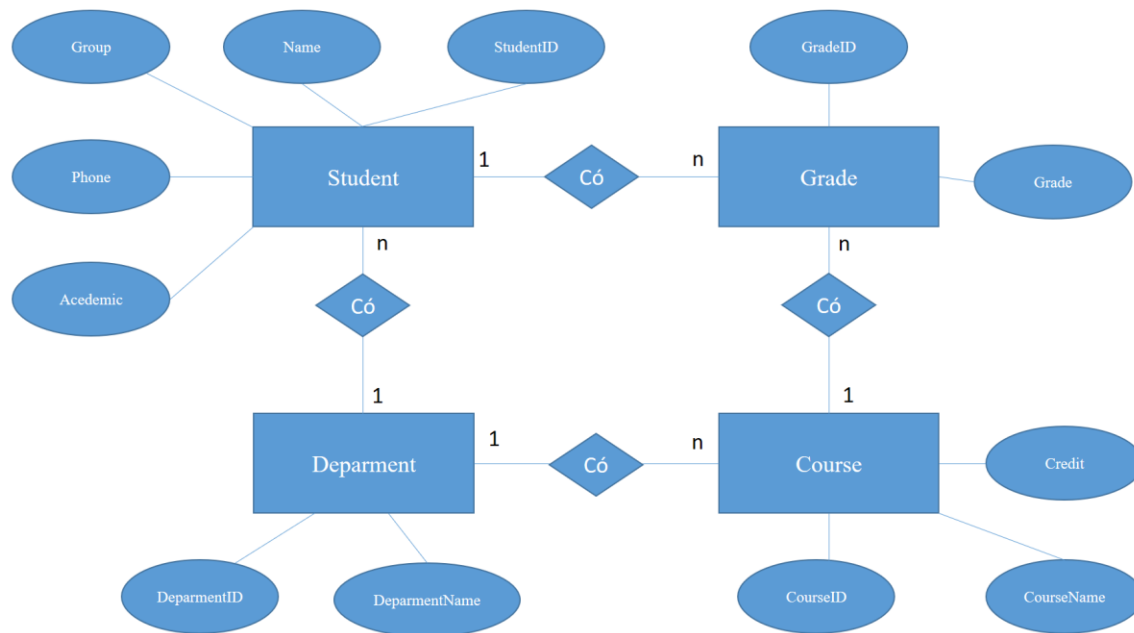
2.1. Mô tả bài toán

Đây là một bài toán xây dựng cơ sở dữ liệu phân tán cho hệ thống “Quản lý sinh viên” nhằm giúp giảm dữ thừa dữ liệu, nhất quán dữ liệu và bảo mật dữ liệu cho hệ thống.

Tại trường đại học X đang đặt ra vấn đề như sau: để tránh xảy ra thất thoát dữ liệu và tăng cường bảo mật, họ muốn chia cơ sở dữ liệu ra hai trạm:

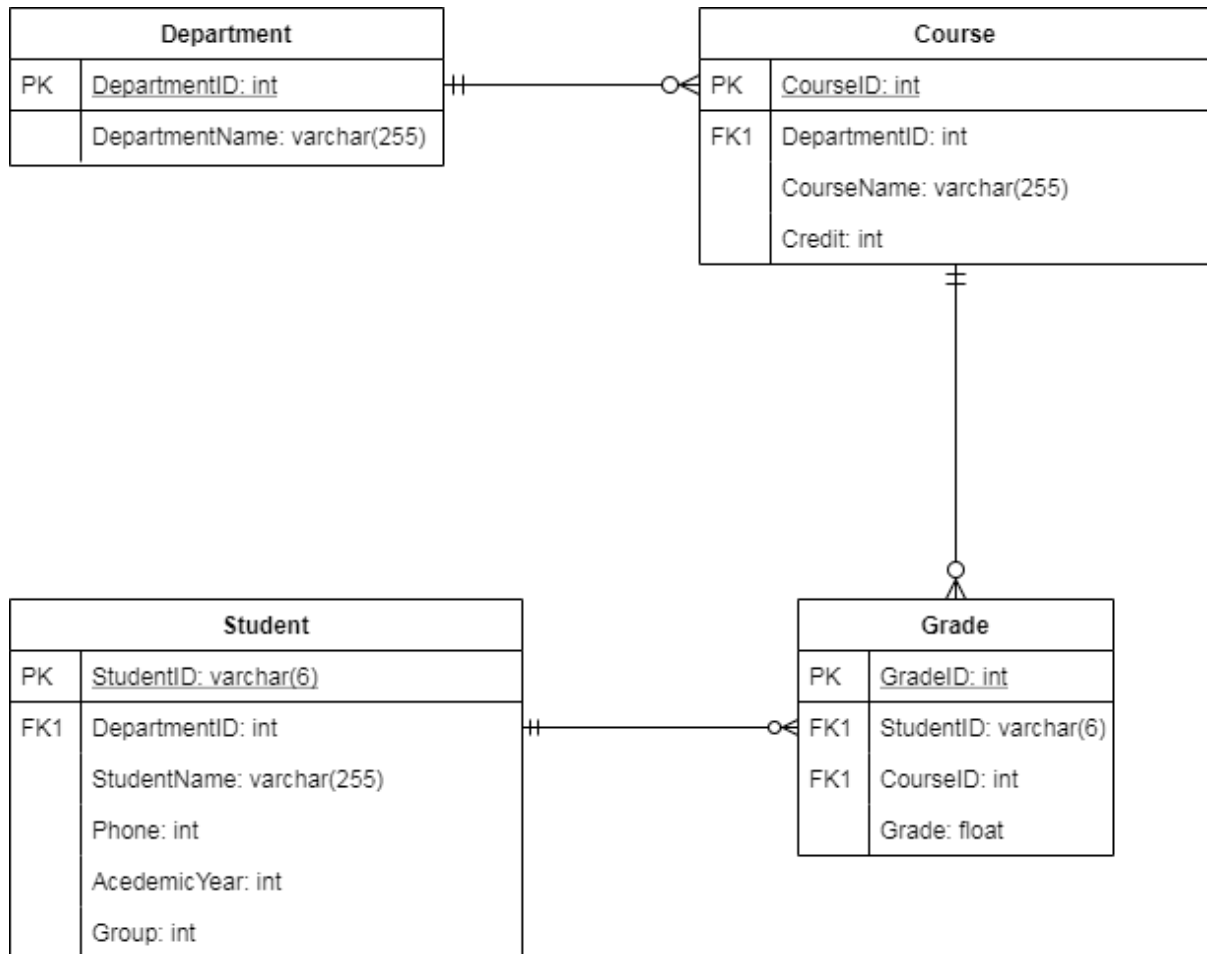
- Tại trạm đầu tiên (nơi trường đại học này được xây dựng) sẽ bao gồm có dữ liệu về:
 - Các thông tin về mã sinh viên, tên tuổi, năm học và nhóm học của sinh viên
 - Điểm số của các sinh viên
 - Thông tin về khoa
 - Thông tin về các khóa học thuộc khoa tự nhiên
- Tại trạm thứ hai (cơ sở hai của trường đại học) sẽ chứa các dữ liệu về:
 - Các thông tin về số điện thoại, khoa theo học của sinh viên
 - Bản dự phòng về điểm số của sinh viên
 - Bản dự phòng về thông tin khoa
 - Thông tin về các khóa học thuộc khoa xã hội

2.2. Mô hình dữ liệu quan hệ



Chương 3: THIẾT KẾ VÀ CÀI ĐẶT CƠ SỞ DỮ LIỆU PHÂN TÁN

3.1. Thiết kế cơ sở dữ liệu phân tán



3.1.1. Phân tích phân mảnh dọc (thuật toán BEA):

3.1.1.1. Đầu vào

						F				
	c1	c2	c3	c4	c5	S1	S2			MSV
ap1	1	0	1	0	1	6	3			28563
ap2	0	1	0	1	0	7	8			29378

3.1.1.2. Ma trận hấp dẫn

	c1	c2	c3	c4	c5
c1	18	0	9	0	9
c2	0	15	0	15	0
c3	9	0	9	0	0
c4	0	15	0	15	0
c5	9	0	0	0	9

3.1.1.3. Ma trận phân cụm hấp dẫn

	C3	C1	C5	C2	C4
C3	9	9	0	0	0
C1	9	18	9	0	0
C5	0	9	9	0	0
C2	0	0	0	15	15
C4	0	0	0	15	15

3.1.1.4. Kết quả

Bảng mới sẽ là						
		C5	C2	C4	C3	C1
	C5	9	0	0	0	9
	C2	0	15	15	0	0
	C4	0	15	15	0	0
	C3	0	0	0	9	9
	C1	9	0	0	9	18
Hai phân vùng sẽ là:						
1	(C,C3,C5,C1)					
2	(C,C2,C4)					

3.1.2. Phân tích phân mảnh ngang (hội vị từ)

3.1.2.1. Tập hội vị từ:

$$Pr = \{ \text{credit}=1; \text{credit}=2; \text{credit}=3; \text{DepartmentID} < 5 \}$$

3.1.2.2. Các vị từ:

$$M1 = \{ \text{credit}=1 \wedge \text{credit}=2 \wedge \text{credit}=3 \wedge \text{departmentid} \leq 5 \}$$

$$M2 = \{ \text{credit}=1 \wedge \text{credit}=2 \wedge \text{credit}=3 \wedge \text{departmentid} > 5 \}$$

$$M3 = \{ \text{credit}=1 \wedge \text{credit}=2 \wedge \text{credit} \neq 3 \wedge \text{departmentid} \leq 5 \}$$

$$M4 = \{ \text{credit}=1 \wedge \text{credit}=2 \wedge \text{credit} \neq 3 \wedge \text{departmentid} > 5 \}$$

$$M5 = \{ \text{credit}=1 \wedge \text{credit} \neq 2 \wedge \text{credit}=3 \wedge \text{departmentid} \leq 5 \}$$

$$M6 = \{ \text{credit}=1 \wedge \text{credit} \neq 2 \wedge \text{credit}=3 \wedge \text{departmentid} > 5 \}$$

$$M7 = \{ \text{credit} \neq 1 \wedge \text{credit}=2 \wedge \text{credit}=3 \wedge \text{departmentid} \leq 5 \}$$

$$M8 = \{ \text{credit} \neq 1 \wedge \text{credit}=2 \wedge \text{credit}=3 \wedge \text{departmentid} > 5 \}$$

$$M9 = \{ \text{credit}=1 \wedge \text{credit} \neq 2 \wedge \text{credit} \neq 3 \wedge \text{departmentid} \leq 5 \}$$

$$M10 = \{ \text{credit}=1 \wedge \text{credit} \neq 2 \wedge \text{credit} \neq 3 \wedge \text{departmentid} > 5 \}$$

$$M11 = \{ \text{credit} \neq 1 \wedge \text{credit}=2 \wedge \text{credit} \neq 3 \wedge \text{departmentid} \leq 5 \}$$

$$M12 = \{ \text{credit} \neq 1 \wedge \text{credit}=2 \wedge \text{credit} \neq 3 \wedge \text{departmentid} > 5 \}$$

$$M13 = \{ \text{credit} \neq 1 \wedge \text{credit} \neq 2 \wedge \text{credit}=3 \wedge \text{departmentid} \leq 5 \}$$

$$M14 = \{ \text{credit} \neq 1 \wedge \text{credit} \neq 2 \wedge \text{credit}=3 \wedge \text{departmentid} > 5 \}$$

$$M15 = \{ \text{credit} \neq 1 \wedge \text{credit} \neq 2 \wedge \text{credit} \neq 3 \wedge \text{departmentid} \leq 5 \}$$

$$M16 = \{ \text{credit} \neq 1 \wedge \text{credit} \neq 2 \wedge \text{credit} \neq 3 \wedge \text{departmentid} > 5 \}$$

3.1.2.3. Kết quả:

- Những vị từ được không được bôi đậm sẽ bị loại bỏ do không thỏa mãn ý nghĩa logic, những vị từ sau sẽ được lựa chọn:

$$M9 = \{ \text{credit} = 1 \wedge \text{departmentid} \leq 5 \}$$

$$M10 = \{ \text{credit} = 1 \wedge \text{departmentid} > 5 \}$$

$$M11 = \{ \text{credit}=2 \wedge \text{departmentid} \leq 5 \}$$

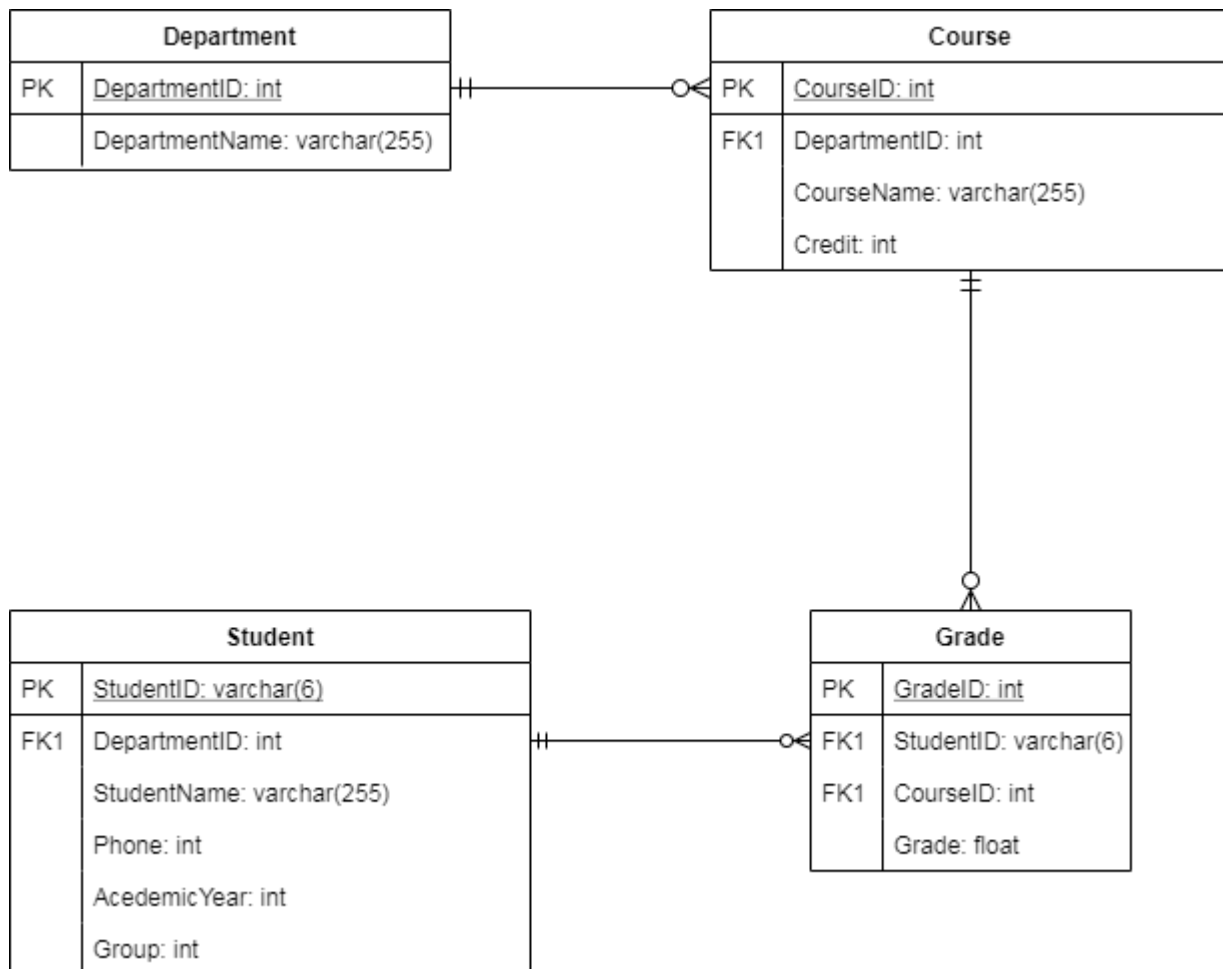
$$M12 = \{ \text{credit}=2 \wedge \text{departmentid} > 5 \}$$

$$M13 = \{ \text{credit}=3 \wedge \text{departmentid} \leq 5 \}$$

$$M14 = \{ \text{credit}=3 \wedge \text{departmentid} > 5 \}$$

3.2. Xây dựng dữ liệu từ điển dữ liệu toàn diện

3.2.1. Global Conceptual Schema:



3.2.2. Data Directory

- Site1 (AP1):
 - o Sitenam: AP1
 - o IP: XXX
- Site2 (AP2):
 - o Sitenam: AP2
 - o IP: XXX

3.2.3. Fragmentation Directory:

3.2.3.1. Phân mảnh ngang:

- Bảng Course được phân mảnh ngang như sau:
 - o Site1: Chứa các môn của mọi khoa bên tự nhiên (departmentID <=5) và có tín chỉ là số dương <=3

- Site2: Chứa các môn của mọi khoa bên xã hội (departmentID >5) và có tín chỉ là số dương <= 3

3.2.3.2. Phân mảnh dọc

- Bảng Student được phân mảnh dọc như sau:
 - Site1: Student(StudentID, StudentName, AcademicYear, Group)
 - Site2: Student(StudentID, Phone, DepartmentID)

3.2.3.3. Replication Directory:

- Các bảng sau được replicated tại 2 site, với dạng Merge Publication (replicate hai chiều):
 - Department
 - Grade

3.2.4. Kiểm soát dữ liệu

3.2.4.1. Student:

- **INSERT/ UPDATE**
 - Lock bảng khi một site đang sử dụng.
 - Site 1: Kiểm soát xem thuộc tính “StudentGroup” của sinh viên có không hợp lệ (> 3 hoặc <=0).
 - Site 2: Kiểm soát xem thuộc tính “DepartmentID” của sinh viên có không hợp lệ.
 - Kiểm soát không bị trùng mã.
- **DELETE:**
 - Kiểm soát xóa sinh viên đúng (sinh viên có tồn tại trong CSDL).
 - Kiểm soát nếu sinh viên có điểm số thì không được xóa.

3.2.4.2. Grade:

- **INSERT/ UPDATE:**
 - Lock bảng khi một site đang sử dụng.
 - Kiểm soát mã sinh viên, mã môn, mã điểm tồn tại.
 - Kiểm soát điểm nằm trong khoảng từ 0-10.
 - Kiểm soát không bị trùng mã.
- **DELETE:**
 - Kiểm soát xóa điểm tồn tại.

3.2.4.3. Course:

- **INSERT/ UPDATE:**
 - Kiểm soát mã khoa.

- Kiểm soát không bị trùng mã.
- Kiểm soát tín chỉ (“Credit”) nằm trong vùng 1-3.
- Lock bảng khi một site đang sử dụng.
- **DELETE:**
 - Kiểm soát xóa môn tồn tại.
 - Kiểm soát không được xóa môn nếu đang có điểm thuộc về môn này.

3.2.4.4. Department:

- **INSERT/ UPDATE:**
 - Lock bảng khi một site đang sử dụng.
- **DELETE:**
 - Kiểm soát không được xóa nếu đang có môn thuộc về khoa này.

3.3. Xây dựng các hàm tính năng phụ trợ:

3.3.1. Tính điểm trung bình của một sinh viên

- Tính năng này cần đưa vào mã sinh viên, sau đó sẽ xuất ra màn hình điểm trung bình (GPA) của sinh viên đó

This student (A00003) has GPA=7.5


3.3.2. Lấy bảng điểm của một sinh viên

- Tính năng này cần đưa vào mã sinh viên, sau đó sẽ xuất ra bảng điểm của sinh viên đó, phục vụ cho việc sinh viên tra điểm của bản thân, cũng như để khi tốt nghiệp có thể in ra bảng điểm giấy.

	GradeID	StudentID	CourseID	Grade
1	1	A00001	1	8.5
2	2	A00001	3	9

3.3.3. Lấy tổng số tín chỉ của một sinh viên

- Tính năng này cần nhập vào mã sinh viên, sau đó sẽ xuất ra tổng số tín chỉ được tích lũy của sinh viên đó, nhằm cho việc sinh viên tra khảo, cũng như để xem điều kiện đi thực tập, cũng như là khoa luận.

 Messages
This student (A00003) has gathered total of 5 credits.