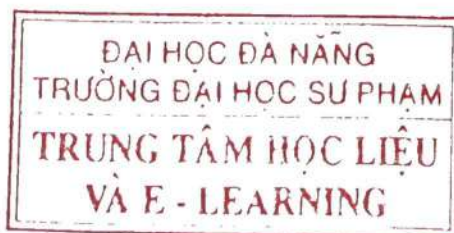


ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC SƯ PHẠM



Nguyễn Trần Quốc Vinh

GIÁO TRÌNH **CƠ SỞ DỮ LIỆU** **NÂNG CAO**



ĐÀ NẴNG – NĂM 2023

LỜI MỞ ĐẦU

Cơ sở dữ liệu là một trong những mảng kiến thức và kỹ năng quan trọng đối với nhân lực làm việc trong ngành công nghệ thông tin, đặc biệt là công nghiệp kỹ thuật phần mềm bởi lẽ cơ sở dữ liệu là một thành phần chính của hệ thống thông tin. Việc cập nhật các hướng phát triển mới vì thế là rất cần thiết đối với người học cũng như nguồn nhân lực CNTT. Đây là động lực chính để tài liệu này được hình thành.

Ngoài các chủ đề truyền thống như i) xử lý văn tin tập trung, ii) giao tác và tương tranh nhằm đảm bảo toàn vẹn dữ liệu và iii) cơ sở dữ liệu phân tán; tài liệu còn trình bày các chủ đề về các phương pháp phân tích và thiết kế cơ sở dữ liệu theo iv) mô hình hoá thông tin theo hướng giao tiếp toàn diện (FCO-IM) và v) thiết kế CSDL thời gian đặt dạng chuẩn 6NF (Anchor modeling), vi) khung nhìn thực, vii) CSDL trong bộ nhớ, viii) CSDL NoSQL và ix) sử dụng các bộ tiêu chuẩn TPC để đánh giá hiệu năng của HQT CSDL.

Mỗi chủ đề được đề cập là một lĩnh vực nghiên cứu và ứng dụng rất rộng. Tài liệu này chủ yếu mang tính giới thiệu các nguyên lý ở mức cơ bản về lý thuyết và thực hành, tạo sự thuận lợi cho người học trong tiếp cận và nâng cao chuyên môn của mình trong các lĩnh vực đó. Mỗi định hướng, mỗi cách tiếp cận mới đều mang đến những ưu điểm mang tính đột phá, nhưng đồng thời mọi phương án tối ưu hoá đều phải trả giá ở những khía cạnh nào đó. Tài liệu này cố gắng phân tích các điểm mấu chốt ở mức cơ bản để người đọc có cái nhìn tổng quan, rõ được nguyên lý hoạt động để đảm bảo các yêu cầu tối thiểu trong quản trị dữ liệu trong khi cung cấp hiệu năng vượt trội; từ đó lựa chọn hướng mới phù hợp nhất, tối ưu cho bài toán mình cần giải quyết trong một lĩnh vực ứng dụng cụ thể.

Dù tác giả đã rất cố gắng trau chuốt tài liệu qua nhiều năm, chắc chắn tài liệu vẫn có thể chứa đựng nhiều thiếu sót. Tác giả rất mong nhận được sự thông cảm, phản hồi về những thiếu sót, chia sẻ những quan điểm để phát triển tài liệu thông qua địa chỉ email (ntquocvinh@{gmail.com, ued.udn.vn}). Nhân đây, tác giả cũng xin chân thành cảm ơn tất cả các chuyên gia đã có những góp ý để hoàn thiện tài liệu và xin cảm ơn bạn đọc đã lựa chọn tài liệu này.

Đà Nẵng, tháng 02 năm 2023
Tác giả

MỤC LỤC

LỜI MỞ ĐẦU.....	i
DANH MỤC VIẾT TẮT.....	vii
DANH MỤC BẢNG BIỂU.....	viii
DANH MỤC HÌNH VẼ	ix
CHƯƠNG 1. TỔNG QUAN.....	1
1.1 CSDL và hệ thống tệp.....	1
1.1.1 Khái niệm CSDL và HQT CSDL	1
1.1.2 Phần cứng lưu trữ.....	2
1.1.3 Hệ thống tệp.....	7
1.2 Tiếp cận quản lý tệp.....	12
1.3 Tiếp cận CSDL và HQT CSDL	13
1.3.1 Một số khái niệm.....	13
1.3.2 Tiếp cận CSDL.....	15
1.3.3 Hệ quản trị CSDL.....	16
1.4 Tiếp cận CSDL định hướng ứng dụng cụ thể	19
CHƯƠNG 2. CƠ SỞ DỮ LIỆU TẬP TRUNG	21
2.1 XỬ LÝ VẤN TIN.....	21
2.1.1 Tổng quan về xử lý truy vấn	21
2.1.2 Chuyển đổi truy vấn sang biểu thức đại số quan hệ.....	23
2.1.3 Tối ưu hoá truy vấn.....	24
2.1.4 Tối ưu hoá trong quá trình viết truy vấn	47
2.2 Giao tác và toàn vẹn dữ liệu	49
2.2.1 Đặt vấn đề	49
2.2.2 Giao tác và phân loại.....	50
2.2.3 Giao tác và tính toàn vẹn của dữ liệu.....	52

2.2.4	Các vấn đề tương tranh	53
2.2.5	Xung đột giữa các giao tác.....	61
2.2.6	Giải quyết vấn đề tương tranh.....	62
2.2.7	Triển khai trong SQL	76
2.2.8	Xử lý tình huống khoá chết.....	82
2.2.9	Hồi phục hệ thống sau sự cố	83
CHƯƠNG 3. CƠ SỞ DỮ LIỆU PHÂN TÁN		87
3.1	HQT CSDL phân tán	87
3.1.1	Khái niệm	87
3.1.2	Kiến trúc	88
3.1.3	Các loại hình CSDL phân tán	89
3.1.4	Ví dụ minh hoạ.....	91
3.2	Lưu trữ dữ liệu trên HQT CSDL phân tán.....	92
3.2.1	Phân mảnh.....	92
3.2.2	Nhân bản	94
3.2.3	Đồng bộ	95
3.2.4	Danh mục phân tán.....	96
3.3	Xử lý truy vấn phân tán	96
3.3.1	Thực thi phép nối	96
3.3.2	Thực thi truy vấn.....	98
3.4	Quản trị giao tác và quá trình đồng bộ hoá.....	98
3.4.1	Giao thức cố định hai pha	99
3.4.2	Bế tắc phân tán	101
3.5	Khởi động lại sau sự cố.....	101
3.6	Giao thức cố định ba pha	102
3.7	Thiết kế cơ sở dữ liệu phân tán.....	103
3.7.1	Các bước thiết kế.....	103

3.7.2	Các chiến lược thiết kế.....	104
CHƯƠNG 4. KHUNG NHÌN THỰC.....		107
4.1	Khái niệm khung nhìn thực.....	107
4.2	Cập nhật khung nhìn thực	108
4.2.1	Các phương pháp cập nhật	108
4.2.2	Biểu diễn truy vấn	110
4.2.3	Cập nhật gia tăng đồng bộ.....	111
4.2.4	Cập nhật gia tăng bất đồng bộ.....	114
4.2.5	Ví dụ tính toán cập nhật gia tăng KNT	117
4.3	Sử dụng khung nhìn thực	121
4.4	“Khung nhìn thực” trong các HQT CSDL không hỗ trợ KNT.....	125
CHƯƠNG 5. FCO-IM & ANCHOR MODELING		130
5.1	Mô hình hoá thông tin theo hướng giao tiếp toàn diện.....	130
5.1.1	Khái niệm	130
5.1.2	Các nguyên tắc cơ bản	132
5.1.3	Mô hình hoá sự giao tiếp.....	136
5.1.4	Ràng buộc.....	149
5.2	Cơ sở dữ liệu thời gian.....	158
5.2.1	Khái niệm	158
5.2.2	CSDL “phi thời gian”	158
5.2.3	Thao tác dữ liệu với CSDL thời gian	160
5.2.4	Truy cập dữ liệu trong CSDL thời gian	160
5.3	Anchor modeling	162
5.3.1	Tổng quan về AM	162
5.3.2	Khái niệm thời gian trong AM.....	163
5.3.3	Mô hình hoá dữ liệu theo AM.....	163
5.3.4	Ví dụ lưu trữ dữ liệu theo AM	165

5.3.5	Truy cập dữ liệu	168
5.3.6	Thao tác dữ liệu.....	170
5.3.7	Công cụ Anchor Modeler.....	170
CHƯƠNG 6. CƠ SỞ DỮ LIỆU TRONG BỘ NHỚ		180
6.1	Khái niệm.....	180
6.2	Tổ chức lưu trữ	181
6.3	Quản trị giao tác.....	185
6.4	Đảm bảo tính lâu bền	185
6.5	Thực thi truy vấn.....	187
6.6	Lựa chọn dữ liệu và thiết bị lưu trữ cho CSDL trong bộ nhớ.....	187
6.7	CSDL NoSQL trong bộ nhớ	189
CHƯƠNG 7. NoSQL		191
7.1	Khái niệm NoSQL	191
7.2	Aggregate và lưu trữ	192
7.3	Họ cột – bảng lớn.....	199
7.4	CSDL đồ thị	200
7.5	Khả năng mở rộng	201
7.5.1	Khái niệm	201
7.5.2	Sharding	202
7.5.3	Nhân bản và đồng bộ	202
7.5.4	Kết hợp sharding và nhân bản.....	203
7.6	Toàn vẹn dữ liệu	203
7.7	Lý do lựa chọn NoSQL	205
CHƯƠNG 8. ĐÁNH GIÁ HIỆU NĂNG.....		209
8.1	Các hệ thống chuẩn TPC	209
8.2	Đơn vị đo	209

8.3 Công tác chuẩn bị..... 211

8.4 Thực hiện đo và báo cáo 212

TỔNG KẾT..... 216

TÀI LIỆU THAM KHẢO 219

CHỈ MỤC..... 222

DANH MỤC VIẾT TẮT

Ký hiệu	Diễn giải
BT	Bài tập
CSDL	Cơ sở dữ liệu
HQT CSDL	Hệ quản trị cơ sở dữ liệu
HTTT	Hệ thống thông tin
KNT	Khung nhìn thực
2PL	Two Phase Lock, khoá chốt hai pha
2PC	Two Phase Commit, cố định hai pha
3PC	Three Phase Commit, cố định ba pha
AM	Anchor Modeling
OLAP	Online Analytical Processing, xử lý phân tích dữ liệu trực tuyến
OLTP	Online Transaction Processing, xử lý giao tác trực tuyến
SQL	Structured Query Language

DANH MỤC BẢNG BIỂU

Bảng 2.1 Ma trận tương dung giữa khoá S và khoá X	63
Bảng 2.2 Ma trận tương dung giữa các khoá có chủ định	72
Bảng 2.3 Mức cô lập và khoá chốt	76
Bảng 2.4 Các mức cô lập giúp giải quyết vấn đề tương tranh.....	77
Bảng 4.1 Dữ liệu bảng KHACH_HANG	119
Bảng 4.2 Dữ liệu bảng MAT_HANG	119
Bảng 4.3 Dữ liệu bảng HD_XUAT	119
Bảng 4.4 Dữ liệu bảng CT_XUAT	119
Bảng 4.5 Dữ liệu KNT.....	120
Bảng 4.6 Ví dụ viết lại truy vấn để sử dụng KNT.....	122
Bảng 5.1 Danh sách các đề tài được đề xuất bởi các giảng viên...	137
Bảng 5.2 Danh sách các đề tài được phân bổ cho sinh viên.....	137
Bảng 5.3 Thông tin về khoảng cách giữa các tỉnh thành theo đường bộ.....	174
Bảng 7.1 Khác nhau cơ bản giữa HQT CSDL quan hệ và Cassandra	199
Bảng 8.1 Các giao tác theo TPC-C và yêu cầu về tỉ lệ.....	210

DANH MỤC HÌNH VẼ

Hình 1.1 Bìa đục lỗ với bảng mã các ký tự	3
<i>Hình 1.2 Bảng đục lỗ 8 cấp (8 lỗ mỗi dòng)</i>	<i>3</i>
Hình 1.3 Bảng từ.....	3
Hình 1.4 Trống từ	3
Hình 1.5 Ổ cứng lưu trữ qua các thời kỳ	5
Hình 1.6 Cấu trúc ổ cứng.....	8
Hình 1.7 Kiến trúc 3 mức theo ANSI/SPARC	16
Hình 1.8 Kiến trúc tiêu biểu theo mô-đun của HQT CSDL	17
Hình 2.1 Các bước thực hiện một câu truy vấn biểu diễn bằng ngôn ngữ bậc cao	22
Hình 2.2 Truy vấn Q trả về danh sách nhân viên sinh sau năm 1977	24
Hình 2.3 Cây truy vấn ban đầu của câu truy vấn SQL Q	25
Hình 2.4 Di chuyển phép chọn xuống phía dưới của cây truy vấn ..	26
Hình 2.5 Áp dụng thêm phép chọn hạn chế trước	27
Hình 2.6 Thay thế phép tích Đề-các và phép chọn bằng phép nối..	27
Hình 2.7 Di chuyển các phép chiếu xuống phía dưới của cây truy vấn	28
Hình 2.8 Hoạt động của trình tối ưu hóa truy vấn	32
Hình 2.9 Tối ưu hoá truy vấn: Điều kiện trong WHERE	47
Hình 2.10 Tối ưu hoá truy vấn: Điều kiện trong HAVING	48

Hình 2.11 Vấn đề mất kết quả cập nhật.....	54
Hình 2.12 Ví dụ vấn đề mất kết quả cập nhật (1).....	54
Hình 2.13 Ví dụ vấn đề mất kết quả cập nhật (2).....	55
Hình 2.14 Vấn đề đọc dữ liệu bản	56
Hình 2.15 Ví dụ vấn đề đọc dữ liệu bản.....	56
Hình 2.16 Vấn đề đọc không lặp lại (1)	57
Hình 2.17 Ví dụ vấn đề đọc không lặp lại (1)	58
Hình 2.18 Vấn đề đọc không lặp lại (2)	58
Hình 2.19 Ví dụ vấn đề đọc không lặp lại (2)	59
Hình 2.20 Vấn đề các phần tử ảo.....	60
Hình 2.21 Ví dụ vấn đề các phần tử ảo.....	61
Hình 2.22 Kịch bản khoá chết	64
Hình 2.23 Giao thức khoá chốt hai pha	65
Hình 2.24 Giải quyết vấn đề mất kết quả cập nhật (1)	66
Hình 2.25 Giải quyết vấn đề mất kết quả cập nhật (2)	67
Hình 2.26 Ví dụ giải quyết vấn đề mất kết quả cập nhật.....	68
Hình 2.27 Giải quyết vấn đề đọc dữ liệu bản	68
Hình 2.28 Giải quyết vấn đề đọc không lặp lại	69
Hình 2.29 Ví dụ giải quyết vấn đề đọc không lặp lại (1)	70
Hình 2.30 Giải quyết vấn đề các phần tử ảo bằng khoá chốt	71
Hình 2.31 Kịch bản giải quyết vấn đề các phần tử ảo bằng khoá chốt đồng bộ	74

Hình 2.32 Ví dụ giải quyết vấn đề các phần tử ảo bằng khoá chốt đồng bộ.....	74
Hình 2.33 Mô phỏng sử dụng mức cô lập READ UNCOMMITTED	79
<i>Hình 2.34 Mô phỏng sử dụng mức cô lập READ COMMITED (a)</i>	79
<i>Hình 2.35 Mô phỏng sử dụng mức cô lập READ COMMITED (b)</i>	80
<i>Hình 2.36 Mô phỏng sử dụng mức cô lập REPEATABLE READ ...</i>	81
<i>Hình 2.37 Mô phỏng sử dụng mức cô lập SERIALIZABLE</i>	82
Hình 3.1 Kiến trúc lược đồ của HQT CSDL phân tán	88
<i>Hình 3.2 Mô hình tối giản của hệ CSDL phân tán</i>	89
<i>Hình 3.3 Hệ CSDL phân tán đồng nhất.....</i>	90
Hình 3.4 Hệ CSDL phân tán bất đồng nhất.....	91
Hình 3.5 Phân mảnh dữ liệu trong CSDL phân tán.....	92
Hình 3.6 Giao thức cố định hai pha (1)	99
Hình 3.7 Giao thức cố định hai pha (2)	100
Hình 3.8 Giao thức cố định 3 pha.....	102
Hình 3.9 Diễn giải giao thức cố định 3 pha.....	103
Hình 4.1 Sơ đồ thực thể - mối quan hệ: quản lý nhập xuất	118
<i>Hình 4.2 Sơ đồ CSDL quan hệ: quản lý nhập xuất</i>	118
<i>Hình 4.3 Truy vấn tính tổng số lượng mỗi mặt hàng đã xuất từ mỗi kho.....</i>	119
Hình 5.1 Minh họa thực hiện phân loại và định danh	144
Hình 5.2 Các ký hiệu đồ họa để biểu diễn IGD.....	146

Hình 5.3 Xây dựng IGD cho loại dữ kiện Sinh viên	147
Hình 5.4 Xây dựng IGD cho loại dữ kiện Đề tài.....	147
Hình 5.5 Xây dựng IGD cho loại dữ kiện Phân bổ	148
Hình 5.6 Xây dựng IGD cho loại dữ kiện Hướng dẫn	148
Hình 5.7 Xây dựng IGD cho loại dữ kiện Mô tả đề tài	149
Hình 5.8 IGD với ràng buộc duy nhất	151
Hình 5.9 IGD với các ràng buộc toàn diện.....	153
Hình 5.10 IGD với ràng buộc con trên một vai trò	154
Hình 5.11 IGD với ràng buộc con trên sự kết hợp của nhiều vai trò	155
Hình 5.12 IGD với ràng buộc loại trừ	156
Hình 5.13 IGD với ràng buộc bản số.....	157
Hình 5.14 Khung nhìn cuối	169
Hình 5.15 Hàm theo thời điểm	169
Hình 5.16 Sơ đồ CSDL được thiết kế dùng công cụ Anchor Modeler	171
Hình 5.17 Sơ đồ tổ chức	175
<i>Hình 5.18 Vé xổ số</i>	176
Hình 5.19 Chuyên ngành học	176
Hình 5.20 Bảng cấp	177
Hình 5.21 Địa chỉ.....	177
Hình 5.22 Kết quả học tập	178
Hình 5.23 Năm sinh và nơi sinh	178

Hình 7.1 Lưu trữ aggregate KHACH_HANG và HD_XUAT.....	193
Hình 7.2 Lưu trữ aggregate KHACH_HANG	194
Hình 7.3 Aggregate chứa thông tin khách hàng và tất cả hoá đơn xuất	195
Hình 7.4 Aggregate chứa thông tin khách hàng	196
Hình 7.5 Aggregate chứa thông tin các hoá đơn xuất	196
Hình 7.6 Mô hình lưu trữ tất cả dữ liệu trong một gói (bucket)....	197
Hình 7.7 Thay đổi thiết kế khoá để phân đoạn dữ liệu trong một gói	198
Hình 7.8 Ví dụ CSDL NoSQL theo mô hình bảng lớn	199
Hình 7.9 Ví dụ CSDL đồ thị.....	200
Hình 8.1 Sơ đồ thực thể - mối quan hệ [29]	211
Hình 8.2 Mô hình hệ thống phục vụ đánh giá hiệu năng của HQT CSDL (tham khảo [29])	212
Hình 8.3 Kết quả đo lường hiệu năng của HQT CSDL theo TPC-C	213