УДК 681.3.016

В.В. КОЛОМЕЙЧУК

РОЗРОБКА ТА ДОСЛІДЖЕННЯ БАЗИ ДАНИХ ДЛЯ СИСТЕМ ОБРОБКИ СТАТИСТИЧНОЇ ІНФОРМАЦІЇ

Abstract: *This article is devoted to problems of development and optimization of specific databases. The analysis of the existing general methods and approaches of database optimization has been given and the need for a qualitatively different approach within the specifics of optimizing statistical information databases has been justified. The implementation of the method of the problem solution has been proposed and illustrated.*
Key words: *database, relational model, database normalization, structure of database.*

Анотація: *Дана стаття присвячена проблемам розробки та оптимізації специфічних баз даних. Проведено аналіз існуючих загальних підходів та методів оптимізації баз даних, обґрунтовано необхідність якісно іншого підходу в рамках специфічної проблематики оптимізації баз даних статистичної інформації. Запропоновано та проілюстровано реалізацію методу рішення поставленої задачі.*
Ключові слова: *база даних, реляційна модель даних, нормальна форма даних, структура бази даних.*

Аннотация: *Данная статья посвящена проблемам разработки и оптимизации специфических баз данных. Проведён анализ существующих общих методов и подходов к оптимизации баз данных, обоснована необходимость качественно иного подхода в рамках специфики проблематики оптимизации баз данных статистической информации. Предложена и проиллюстрирована реализация метода решения поставленной задачи.*
Ключевые слова: *база данных, реляционная модель данных, нормальная форма данных, структура базы данных.*

1. Вступ

Перехід від роботи комп'ютерних систем тільки з короткостроковою пам'яттю до використання короткострокової та довгострокової пам'яті одночасно дозволив накопичувати в системі як програмні дані, так і інформаційні. Тобто, для того, щоб виконати обробку деяких даних, більше не потрібно кожен раз програмувати апаратуру та вводити необхідну інформацію для обробки. Тепер можна ввести один раз і зберігати в пам'яті комп'ютерної системи набір програмних та інформаційних даних, після чого лише обирати, яку інформацію та яким чином необхідно обробити. Це призвело до значної економії часу і роботи операторів комп'ютерних систем.

Поява апаратного забезпечення для довгострокового збереження створила цілу нову галузь у комп'ютерних науках. З одного боку, почався розвиток апаратних рішень, програмного забезпечення низького рівня та алгоритмів шифрування й кодування, що забезпечило швидкий розвиток засобів накопичення інформації. З іншого боку, це спричинило поштовх у розробці програмного забезпечення високого рівня, яке реалізовувало людино-машинний і програмно-програмний інтерфейс. Однією із головних проблем програмного забезпечення високо рівня була і досі залишається реалізація взаємодії з незалежним програмним забезпеченням обробки інформації для роботи з великими об'ємами інформації. Рішення цієї проблеми беруть на себе системи управління базами даних (СУБД). Вони реалізують інтерфейс між апаратними засобами накопичення інформації та програмним забезпеченням обробки інформації. Такий рівень незалежності є необхідним, оскільки розробник програмного забезпечення обробки інформації має бути незалежним від змін апаратних засобів накопичення інформації та реалізації таких проблем збереження інформації, як цілісність чи безпека даних.

Сучасні системи управління базами даних надають додатковий рівень незалежності: вони реалізують функціонал роботи з базою даних, не вводячи жорстких обмежень на логічну структуру

ISSN 1028-9763. Математичні машини і системи, 2009, № 4

самої бази даних. У даній статті піде мова саме про розробку та оптимізацію баз даних для специфічних задач.

2. Методи та підходи до проблеми оптимізації баз даних

У більшості випадків під оптимізацією бази даних розуміють оптимізацію роботи з базою даних, оскільки саме це найбільш є важливим для кінцевого програмного продукту, що виконує обробку даних. В оптимізації роботи з базою даних виділяють три напрями: оптимізація файлових серверів та апаратних засобів, оптимізація запитів до бази даних та оптимізація структури бази даних [1].

Оптимізація роботи файлових серверів має найменший вплив на оптимізацію бази даних. Звичайно, під оптимізацією серверів розуміють удосконалення та налагодження апаратних і програмних засобів та операційної системи сервера. Оптимізація запитів, напроти, більше відноситься до програмного забезпечення обробки інформації. За допомогою запитів програмний продукт з бази даних отримує необхідні дані. Оптимізація запитів призводить до підвищення швидкості взаємодії між програмним продуктом та базою даних і зменшує збитковість переданих через лінії зв'язку даних. Ці два напрями є важливими в оптимізації роботи з базами даних.

Під оптимізацією бази даних, перш за все, мається на увазі оптимізація структури бази даних, організація інформації в базі даних на логічному рівні. На сьогоднішній день найбільш розповсюдженими є реляційні бази даних, розроблені на основі реляційної моделі даних, створеної доктором Е. Коддом. Теорія реляційної моделі даних описує математичну й логічну модель щодо організації та роботи з інформацією, що зберігається в базах даних [2]. До переваг цієї теорії можна віднести строгу логічну формалізацію, незалежність даних та можливість оптимізації структури бази даних за допомогою нормалізації. Під нормалізацією розуміється приведення інформації в базі даних до нормальних форм, що дозволяє звільнитися від збитковості та забезпечити цілісність даних. Усього на сьогоднішній день виділяють вісім нормальних форм [3]. Кожна наступна нормальна форма вимагає виконання умов попередньої нормальної форми.

Перша нормальна форма вимагає від організації таблиці, щоб кожен атрибут таблиці був атомарним, тобто атрибут має складатися з єдиного значення і не може містити в собі переліку значень. Таблиця знаходиться в другій нормальній формі, якщо вона є в першій нормальній формі та всі атрибути таблиці знаходяться у повній залежності від ключового атрибута, а не тільки від деякої його частини. До третьої нормальної форми відносяться такі таблиці другої нормальної форми, в яких неключові атрибути залежать тільки від ключового атрибута. Існує також модифікація третьої нормальної форми, яка називається нормальною формою Бойса-Кодда. Таблиця знаходиться в нормальній формі Бойса-Кодда, якщо вона знаходиться в третій нормальній формі та ключовий атрибут не має функціональних залежностей від неключових атрибутів [3]. Таблиця знаходиться в четвертій нормальній формі, якщо вона знаходиться в нормальній формі Бойса-Кодда і всі многозначні залежності є функціональними. Таблиця четвертої нормальної форми знаходиться в п'ятій нормальній формі, якщо у своїй структурі містить обмеження на допустимі комбінації атрибутів. П'ята нормальна форма рідко використовується на практиці і є предметом наукових досліджень. Таблиця знаходиться в доменно-ключовій нормальній формі, якщо забезпечується умова цілісності даних при будь-яких змінах в атрибутах таблиці. Таблиця

знаходиться в шостій нормальній формі, якщо виконуються умови п'ятої нормальної форми та усі залежності є функціональними. Інколи шосту нормальну форму асоціюють з доменно-ключовою нормальною формою [2] .

Завдяки розвинутому математичному та логічному апарату, теорія реляційної моделі даних дозволяє проводити оптимізацію бази даних на логічному рівні, тобто на рівні організації структури бази даних. Описані вище форми нормалізації дозволяють знизити вірогідність появи аномалій даних та збитковість даних, впровадити в саму структуру бази даних реалізацію методів підтримки цілісності даних, зробити базу даних зрозумілою для роботи як людині, так і програмному забезпеченню. Але даний підхід має один значний недолік усіх стандартизованих методів: специфіка конкретних задач не враховується, в результаті чого залишається місце для вдосконалень.

3. Аналіз предметної області та обґрунтування необхідності специфічного підходу до розробки бази даних статистичної інформації

Одними з найбільших баз даних, що існують нині, є бази даних (БД) статистичної інформації. Вони широко використовуються в дослідницькій діяльності, державних установах, бізнесі та багатьох інших сферах. Сучасні системи прогнозування та аналізу також потребують великих об'ємів статистичних даних. На державному рівні навіть створюються архівні бази даних статистичної інформації для накопичення та можливості подальшої обробки. Характерними рисами цих баз даних є великі об'єми інформації, що носять архівний характер, тобто майже не змінюються: нестандартизованість структур таблиць, оскільки найчастіше нова статистична інформація організовується в таблиці, прості для відображення, але складні для їх комп'ютерної обробки; низький рівень зв'язності даних між таблицями, викликаний тим, що таблиці статистики рідко пов'язані між собою, а усі близькі дані намагаються розмістити в одній таблиці.

Такі особливості баз даних статистичної інформації породжують ряд специфічних проблем: база даних складається з великої кількості таблиць. При роботі з базою даних програмне забезпечення має знати структуру конкретної таблиці, що при відсутності стандартизації структур викликає складності та ріст збитковості даних, що передаються по лініях зв'язку [4]; приведення таблиць до нормальних форм реляційної моделі призводить до розбиття однієї логічної таблиці на ряд взаємопов'язаних таблиць бази даних. Велика кількість таблиць бази даних призводить до збільшення часу пошуку, ускладнення запитів до бази даних. Операторам стає складніше відслідковувати й працювати з такою базою даних. Розробка програмного забезпечення теж стикається зі специфічними проблемами. Реалізація інтерфейсу вигляду даних на логічному рівні між системою управління базою даних та програмним рішенням стає задачею розробника прикладного програмного забезпечення, що, у свою чергу, нехтує принципом незалежності реалізації бази даних від реалізації програмного забезпечення, яке працює з базою даних [5]. Загальна складність реалізації як на стороні бази даних, так і на стороні програмного забезпечення підвищують вірогідність появи помилок при роботі з такою системою [4] .

На фоні цих проблем стає очевидною необхідність якісно іншого підходу до розробки та оптимізації баз даних статистичної інформації. Потужний механізм нормалізації реляційних таблиць

дає ряд важливих переваг, але стандартизований підхід не враховує специфіки баз даних обліку статистичної інформації і не звільняє від ряду специфічних для даних баз даних недоліків.

4. Оптимізація баз даних статистичних даних на логічному рівні

Під логічним рівнем оптимізації бази даних маємо на увазі оптимізацію структури бази даних та виділення логічних одиниць (таблиць) у базі даних. Оскільки база даних статистичної інформації складається з набору таблиць статистичних даних, то вони і є логічною одиницею такої бази даних (табл. 1). Структурно база даних статистичної інформації складається з ряду таблиць статистичних даних та декількох службових таблиць, в яких зберігаються опис та зв'язки між таблицями статистичних даних.

Таблиця 1. Вихідна таблиця статистичних даних. Кількість суб'єктів ЄДРПОУ за галузями економіки станом на 1 квітня 2009 року

	Код секції КВЕД	Суб'єктів ЄДРПОУ			
		Усього	У % до загальної кількості	У тому числі	
				із статусом юридичної особи	без статусу юридичної особи
Усього		1235183	×	1175683	59500
У тому числі					
Сільське господарство, мисливство, лісове господарство	A	85973	7,0	84494	1479
Рибальство, рибництво	B	1976	0,2	1924	52
Промисловість		124626	10,1	118443	6183
Добувна промисловість	C	4844	0,4	4463	381
Переробна промисловість	D	114214	9,2	109569	4645
Виробництво і розподілення електроенергії, газу та води	E	5568	0,5	4411	1157
Будівництво	F	90118	7,3	87294	2824

Першим кроком в оптимізації бази даних статистичної інформації є нормалізація вихідних таблиць статистичних даних та службових таблиць (табл. 2). Це дозволяє зменшити збитковість даних та покращити показники однозначності й цілісності інформації. Для цього ми використовуємо другу та третю нормальні форми або нормальну форму Бойса-Кодда. Наступним кроком є створення універсального інтерфейсу для розробників програмного забезпечення, що дозволить звільнитися від необхідності програмному рішенню знати структуру бази даних. Для реалізації цього кроку існують два підходи: стандартизація структури бази даних та таблиць бази даних; надання програмному забезпеченню даних про структуру таблиці бази даних разом із самою таблицею. Для баз даних обліку статистичних даних пропонується використовувати комбінований підхід, а саме перехід від збереження в базі даних таблиць статистичних даних до збереження опису таблиць статистичних даних.

Таблиця 2. Нормалізована таблиця статистичних даних

Галузь економіки	Код секції КВЕД	Усього суб'єктів ЄДРПОУ	Суб'єктів ЄДРПОУ у % до загальної кількості	Суб'єктів ЄДРПОУ, у тому числі із статусом юридичної особи	Суб'єктів ЄДРПОУ, у тому числі без статусу юридичної особи
Сільське господарство, мисливство, лісове господарство	A	85973	7	84494	1479
Рибальство, рибництво	B	1976	0,2	1924	52
Промисловість		124626	10,1	118443	6183
Добувна промисловість	C	4844	0,4	4463	381
Переробна промисловість	D	114214	9,2	109569	4645
Виробництво та розподілення електроенергії, газу та води	E	5568	0,5	4411	1157
Будівництво	F	90118	7,3	87294	2824
Усього		1235183	×	1175683	59500

Цей підхід надає ряд переваг. Він дозволяє стандартизувати структуру самої бази даних, тобто різні бази даних будуть мати одну структуру, що значно полегшить роботу з розподіленими базами даних та спростить злиття декількох баз даних. Крім того, база даних буде зберігати та надавати програмному забезпеченню інформацію не тільки про кінцеві значення таблиці, а й опис структури самої таблиці статистичних даних, що підвищує гнучкість як бази даних, так і програмного забезпечення. Також зникає необхідність у створенні службових таблиць, оскільки вся інформація опису може знаходитися в одній таблиці бази даних разом з описом структури таблиці статистичних даних.

Суть запропонованого підходу: база даних має складатися з чотирьох службових таблиць опису таблиць статистичної інформації. Першою є таблиця опису таблиць статистичних даних. Вона в собі містить назви таблиць статистичних даних та усі службово-довідкові дані, наприклад, дату створення, термін актуальності, регіон та ін. Друга таблиця бази даних містить у собі опис колонок таблиць статистичної інформації. Серед атрибутів другої таблиці обов'язково має бути посилання на першу таблицю, оскільки таким чином вказується належність конкретної колонки до конкретної таблиці статистичної інформації. Як і перша, друга таблиця може містити в собі атрибути з додатковими службовими даними, наприклад, одиницями виміру. Третя таблиця бази даних подібна до другої, але містить у собі опис рядків таблиць статистичної інформації, а не колонок. Атрибут, що вказує на належність певного рядка певній таблиці, також є обов'язковим. Четверта таблиця бази даних містить у собі кінцеві значення таблиць статистичних даних. Серед атрибутів четвертої таблиці є обов'язковими атрибути посилання на другу та третю таблиці бази даних. Таким чином, певне значення знаходиться на перетині певних колонки та рядка певної таблиці статистичних даних. Це можливо завдяки попередній нормалізації таблиць статистичних даних. Отримані чотири таблиці бази даних повинні знаходитися в нормальній формі Бойса-Кодда (табл. 3). Також серед атрибутів четвертої таблиці окреме місце займає атрибут посилання на

першу таблицю. Наявність атрибута належності певного значення певній таблиці статистичних даних підвищує цілісність даних бази даних, оскільки може бути виконана перевірка належності певного значення, його колонки та рядка до певної таблиці статистичних даних. Така перевірка допомагає виявити аномалії даних, коли значення, його колонка або його рядок не належить таблиці статистичних даних, якій належать два інших значення.

Таблиця 3. Приведення таблиці статистичних даних до запропонованого вигляду

Номер таблиці	Назва таблиці	Рік створення
1	Кількість суб'єктів ЄДРПОУ за галузями економіки станом на 1 квітня 2009 року	2009
…	…	…

Номер колонки	Номер таблиці	Назва колонки	Розмір-ність
1	1	Галузь економіки	б\р
2	1	Код секції КВЕД	б\р
3	1	Усього суб'єктів ЄДРПОУ	б\р
4	1	Суб'єктів ЄДРПОУ у % до загальної кількості	б\р
5	1	Суб'єктів ЄДРПОУ, у тому числі із статусом юридичної особи	б\р
6	1	Суб'єктів ЄДРПОУ, у тому числі без статусу юридичної особи	б\р
…	…	…	…

Номер рядка	Номер таблиці	Назва рядка
1	1	Сільське господарство, мисливство, лісове господарство
2	1	Рибальство, рибництво
3	1	Промисловість
4	1	Добувна промисловість
5	1	Переробна промисловість
6	1	Виробництво та розподілення електроенергії, газу та води
7	1	Будівництво
8	1	Усього
…	…	…

Номер значення	Номер таблиці	Номер колонки	Номер рядка	Розмірність	Значення
1	1	1	1	б\р	A
2	1	1	2	б\р	B
3	1	1	3	б\р	
4	1	1	4	б\р	C
5	1	1	5	б\р	D
6	1	1	6	б\р	E
7	1	1	7	б\р	F
8	1	1	8	б\р	
9	1	2	1	б\р	85973
10	1	2	2	б\р	1976
11	1	2	3	б\р	124626
12	1	2	4	б\р	4844
13	1	2	5	б\р	114214
14	1	2	6	б\р	5568
15	1	2	7	б\р	90118
16	1	2	8	б\р	1235183
17	1	3	1	б\р	7
18	1	3	2	б\р	0,2
19	1	3	3	б\р	10,1
20	1	3	4	б\р	0,4
21	1	3	5	б\р	9,2
22	1	3	6	б\р	0,5
23	1	3	7	б\р	7,3
24	1	3	8	б\р	x
25	1	4	1	б\р	84494
26	1	4	2	б\р	1924
27	1	4	3	б\р	118443
28	1	4	4	б\р	4463
29	1	4	5	б\р	109569
30	1	4	6	б\р	4411
31	1	4	7	б\р	87294
32	1	4	8	б\р	1175683
33	1	5	1	б\р	1479
34	1	5	2	б\р	52
35	1	5	3	б\р	6183
36	1	5	4	б\р	381
37	1	5	5	б\р	4645
38	1	5	6	б\р	1157
39	1	5	7	б\р	2824
40	1	5	8	б\р	59500
…	…	…	…	…	…

До переваг бази даних статистичних даних, приведеної до такого вигляду, можна віднести забезпечення цілісності та однозначності даних, стандартизацію структури бази даних та полегшення виконання пошуку.

5. Висновок

У даній статті була розглянута проблематика розробки та оптимізації специфічних баз даних. У той час, як розвиваються теоретично-математичні апарати моделювання баз даних, поряд із загальними проблемами побудови оптимізованих баз даних виникають специфічні проблеми, тісно пов'язані зі специфікою побудови та використання окремих баз даних.

Як приклад для ілюстрації було використано базу даних статистичної інформації, оскільки такі бази даних є широко розповсюдженими і мають певну специфіку розробки й використання. В ході роботи над розробкою оптимізованої бази даних статистичної інформації були використані сучасні підходи, основані на теорії реляційної моделі даних. Оптимізація була проведена як за допомогою загальних методів нормалізації реляційних баз даних, так і за допомогою певного методу, розробленого для баз даних статистичної інформації. Окремо мають бути проведені дослідження поведінки оптимізованої бази даних у різних умовах реального функціонування та

написання програмної оболонки для розвитку і впровадження розробленого стандартизованого підходу й полегшення роботи розробників програмного забезпечення обробки статистичної інформації.

СПИСОК ЛІТЕРАТУРИ

1. Диго С.М. Проектирование и использование баз данных. – Москва: Финансы и статистика, 1995. – 592 с.
2. Мейер М. Теория реляционных баз данных. – Москва: Мир, 1998. – 608 с.
3. Джексон Г. Проектирование реляционных баз данных для использования с микро-ЭВМ. – Москва: Мир, 2001. – 256 с.
4. Шумаков П.В. Delphi 3.0 и создание баз данных. – Москва: Нолидж, 1997. – 320 с.
5. Larson B. Delivering Business Intelligence with Microsoft SQL Server 2008. – McGraw-Hill Osborne Media; 2 edition. – 792 p.

Стаття надійшла до редакції 21.04.2009