УДК 681.3.016.001.57

ОПТИМІЗАЦІЯ РОЗПОДІЛУ ДАНИХ В КОРПОРАТИВНІЙ ІНФОРМАЦІЙНІЙ СИСТЕМІ

Л.А. Павленко
(Харківський національний економічний університет)

Наведено методи оптимізації розподілу даних в корпоративній інформаційній системі, які реалізують наступні критерії: мінімізація дублювання даних; максимально можливе наближення даних до місць їхнього використання; мінімізація часу відгуку системи в додатках.

база даних, система баз даних, розподілена інформаційна система, критерії розподілу даних, методи оптимізації

Постановка проблеми. Концепції локалізації і глобалізації бізнесу, нового системного проектування, реінжінірінга, створення кіберкорпорацій, організації єдиного інформаційного простору, єдиного системного образу інформаційної системи (ІС) придбали властивість імператива інформаційних технологій і піднімають на новий рівень проблеми організації інформаційних ресурсів та інструментальних засобів розробки і підтримки баз даних (БД) розподілених інформаційних систем (РІС). База даних розподіленої корпоративної інформаційної системи – це складний комплекс мультибаз даних у гетерогенному мережному середовищі, який включає як деталізовані транзакційні сукупності даних — розподілені (дистрибутивні) БД та системи баз даних з віддаленим доступом, так і сукупності агрегованих даних, призначених для аналітичної обробки – сховища і вітрини даних. Більш того, діапазон транзакційних БД корпоративної ІС надзвичайно широкий: від успадкованих систем із плоскими файлами даних до ”БД, які кочують” мобільних ІС. Сховища даних теж не обов'язково являють собою монолітну архітектуру, а теж можуть бути розподілені [1 – 4].

У загальному випадку розподілена або дистрибутивна база даних (РБД) – сукупність безлічі взаємозалежних баз даних, розподілених у комп'ютерній мережі. Така БД може бути представлена або, як система баз даних з віддаленим доступом на основі розширеної архітектури клієнт-сервер, або як РБД, що відповідає фундаментальному принципу РБД "Rule Zero" [5]. У першому випадку дані зберігаються на серверах, додатки виконуються на клієнтських робочих станціях, причому для клієнтів

відомі місця локалізації даних у мережі. В другому випадку виконується правило: "Для користувачів розподілена БД є нерозподіленою", тобто всі проблеми організації даних не відносяться на користувача системи. Дж. Дейт [5] сформулював дванадцять цілей забезпечення цього принципу. В ідеальному випадку система управління розподіленою базою даних (СУРБД) повинна забезпечувати прозорість розподілу даних.

При проектуванні корпоративної БД або бази даних розподіленої ІС виконується вибір варіанта організації збереження, доступу до даних і обробки даних: централізована БД, розподілена БД (РБД), розподілена обробка. Крім того, виконується поділ БД на транзакційну й аналітичну частини. На рис. 1 приведена схема стратегій організації даних у розподілених інформаційних системах. Централізована стратегія пов'язана із розміщенням БД на єдиному сервері та організації доступу клієнтів до даних.

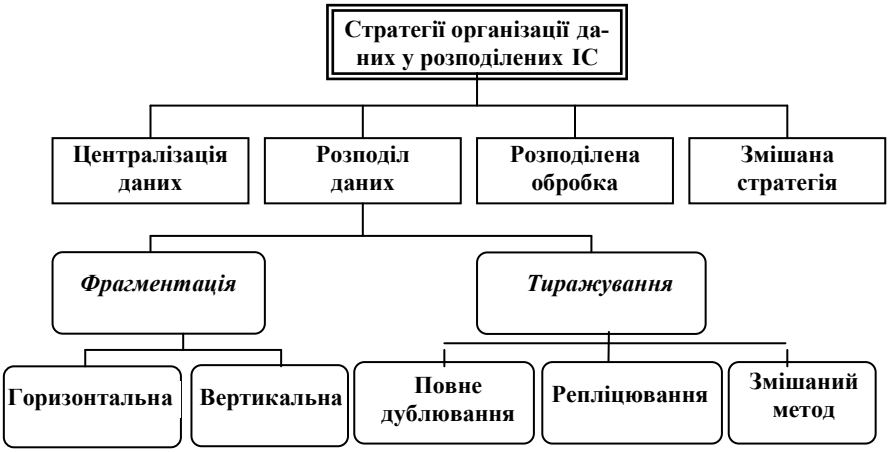

Рис. 1. Стратегії організації даних у розподілених ІС

Архітектура такої обчислювальної системи може бути або файл-серверна, або клієнт-серверна. Стратегія розподілу БД поділяється на фрагментацію та тиражуання. В першому випадку окремі фрагменти БД не перетинаються та повинні бути розташовані на окремих серверах. Тут розрізнюють горизонтальну та вертикальну фрагментації. Стратегія тиражування поділяється наступним чином: метод повного дублювання – коли кожен сервер зберігає усю БД, метод виконання реплік – копій окремих сегментів БД на визначених серверах у мережі та змішаний метод – частке дублювання та часткове репліцювання. Розподілена обробка зв'язана з розподілом загальної задачі обчислень на декілька ком-

п'ютерів і зв'язана з архітектурою обчислень клієнт-сервер. Змішана стратегія має розповсюдження у корпоративних ІС з гетерогенним обчислювальним середовищем та вміщує в собі усі або частково усі перелічені стратегії.

Переваги розподілу даних полягають у полегшенні забезпечення паралелізму обробки, незалежності, гнучкості, доступності даних для кінцевих користувачів. Недоліки полягають у необхідності забезпечення безпеки, надійності, вирішення проблеми „поширення відновлення даних", забезпечення семантичної цілісності даних, складністю та вартістю розробки та підтримки [5, 6].

Аналіз останніх досліджень і публікацій. Проблема проектування складної автоматизованої інформаційної системи і бази даних продовжує залишатися актуальною, незважаючи на досягнуте країнами ЄС у 70-х роках 20 сторіччя угоди про створення "Еврометода", що базується на успішно апробованих методологіях: SSADM – Великобританія, Merise – Франція, IDEF – США, Dafne – Італія, NIAM – Нідерланди та ін.

Формулювання цілей статті. Метою роботи є розробка методів та моделей оптимізації розподілу даних в корпоративній інформаційній системі та підтримки єдиного інформаційного простору територіально розгалуженої організації, заснованих на виконанні наступних критеріїв розподілу: мінімізація дублювання даних; максимально можливе наближення даних до місць їхнього використання; мінімізація часу відгуку системи в додатках.

Виклад основного матеріалу. При проектуванні однорідної РБД із глобальною схемою або розподіленої БД, яка підтримується в однорідному комп'ютерному середовищі (на однакових апаратній і програмній платформах засобами однієї СУРБД) доцільно дотримуватися наступних етапів проектування [2, 3]:

1) формулювання вимог до розроблювальної БД;

2) концептуальне інфологічне проектування або проектування суперсхеми даних, які підлягають збереженню, у вигляді і ER-діаграми, що відповідає варіантові нерозподіленої БД;

3) вибір технічного і програмного видів забезпечення і СУРБД;

4) концептуальне даталогічне проектування або проектування даталогічної моделі, що відповідає суперсхемі інфологічної моделі. Нормалізація логічного подання даних. Одержання реляційної моделі;

5) виконання етапів 1 – 4 для кожного з вузлів мережі – проектування глобальних ER-діаграм і даталогічних моделей;

6) аналіз отриманих структур даних – пошук загальних для різних вузлів мережі масивів і елементів даних;

7) розподіл загальних (для різних вузлів мережі) масивів довідників даних. Рішення питань про фрагментацію і реплікацію цих масивів;

8) розподіл загальних для різних вузлів мережі оперативних даних і даних перетинання;

9) оптимізація запитів – вибір вузлів мережі, де будуть виконуватися обчислення. Розподіл загальних для різних вузлів мережі даних, що обчислюються;

10) проектування допоміжних масивів для передачі даних по мережі;

11) розподіл і розміщення загального для всієї розподіленої інформаційної системи системного каталогу – репозитарія;

12) фізичне проектування фрагментів РБД і допоміжних масивів для кожного вузла мережі.

При виконанні п. 7 – 11 використовуються наступні критерії розподілу даних: максимально можливе наближення даних до місць їхнього використання; мінімізація дублювання даних; мінімізація часу відгуку системи в додатках; мінімізація мережного трафіка.

Рішення задачі розподілу даних виконується при наступних умовах:

1) БД має глобальну схему;

2) усі вузли мережі, де будуть розташовані дані, об'єднані єдиною мережею;

3) трафік для всіх сегментів мережі однаковий;

4) пропускна здатність мережі досить висока для забезпечення достатньої швидкості одержання відповіді на запити користувачів.

Рішення задач розподілу даних виконується на підставі відомих методів оптимізації [7].

Перша з задач розподілу – фрагментація даних єдиного інформаційного простору корпорації формулюється наступним чином.

Існують N фрагментів даних f_i, таких що не перетинаються, та N серверів S_j у мережі.

Необхідно розмістити фрагменти так, щоб кожний з них знаходився на окремому сервері і загальна вартість F устаткування системи була б мінімальною.

Нехай існує оцінка об’єму кожного фрагмента f_i, яка зіставляється із розміром пам’яті кожного комп'ютера S_j. Позначимо C_{ij} оцінку значення вартості розміщення фрагмента f_i на вузлі S_j, яка викликана витратами на модернізацію устаткування у зв'язку з необхідністю розміщення там визначеної порції даних. Тоді функція мети має вигляд:

$$F = \sum_{i=1}^{N} \sum_{j=1}^{N} x_{ij} \cdot c_{ij} \longrightarrow \min$$

при умовах:

$$\sum_{j=1}^{N} x_{ij} = 1;\ \sum_{i=1}^{N} x_{ij} = 1;\ x_{ij} \geq 0 \text{ та мають бути цілими, } i \in [1, N],\ j \in [1, N],$$

де x_{ij} – план розміщення фрагмента f_i на сервері S_j.

Умови (2) гарантують розміщення кожного фрагмента на окремому вузлі.

Безсумнівною перевагою стратерії фрагментації є відсутність необхідності вирішувати задачу «поширення відновлення даних» [5]. Недоліком є низький ступінь локалізації даних у запитах, які є адресованими різним серверам. Взагалі фрагментація є ідеалізованою і такою стратегією, що є рідко затребуваною.

Найбільш розповсюдженою є стратегія тиражування даних. Рішення цієї задачі пропонується в наступній постановці.

Мається N однакових для різних вузлів у мережі компонентів даних f_i і M вузлів S_j.

Позначимо C_{ijk} ступінь локалізації і-го масиву на j-м комп'ютері при реалізації k-го запиту до даних.

$$C_{ijk} = \frac{V_{ijk}}{\sum_{i=1}^{M} \sum_{j=1}^{N} V_{ijk}},$$

де V_{ijk} – об’єм даних і-го масиву, необхідних на j-м комп'ютері при реалізації k-го запиту; $\sum_{i=1}^{M} \sum_{j=1}^{N} V_{ijk}$ – загальний об’єм даних, необхідних при реалізації k-го запиту, який звертається до всіх масивів, розташованих на всіх комп'ютерах у мережі.

Тоді узагальнена ступінь локалізації масиву f_i на комп'ютері S_j по всіх запитах може бути подана в такий спосіб

$$C_{ij} = \sum_{k=1}^{K} C_{ijk}.$$

Значення C_{ij} оцінюється аналітиками, що проектують додатки, в яких реалізуються регламентовані запити до даних.

Оптимізаційна задача розподілу даних формулюється в такий спосіб. Необхідно розподілити дані між вузлами збереження так, щоб загальний ступінь локалізації даних F у додатках, які виконуються у мережі, був би максимальним.

$$F = \sum_{i=1}^{M}\sum_{j=1}^{N} c_{ij} \cdot x_{ij} \longrightarrow \max ,$$

де x_{ij} – об'єм даних масиву i на сервері j при наступних обмеженнях: $\sum_{j=1}^{N} x_{ij} = U_i$ – об'єм i-го компонента; $\sum_{i=1}^{M} x_{ij} = Q_j$ – об'єм даних, які можуть бути розміщеними на вузлі j; $\sum_{i=1}^{M} U_i = \sum_{j=1}^{N} Q_j$; $x_{ij} \geq 0$; $i \in [1, M]$; $j \in [1, N]$.

На рис. 2 представлена схема розподілу однакових компонентів бази даних A, B, C, D на серверах S_1, S_2, S_3, S_4, S_5 у пропорції x_{ij} (i – номер компонента, j – номер сервера).

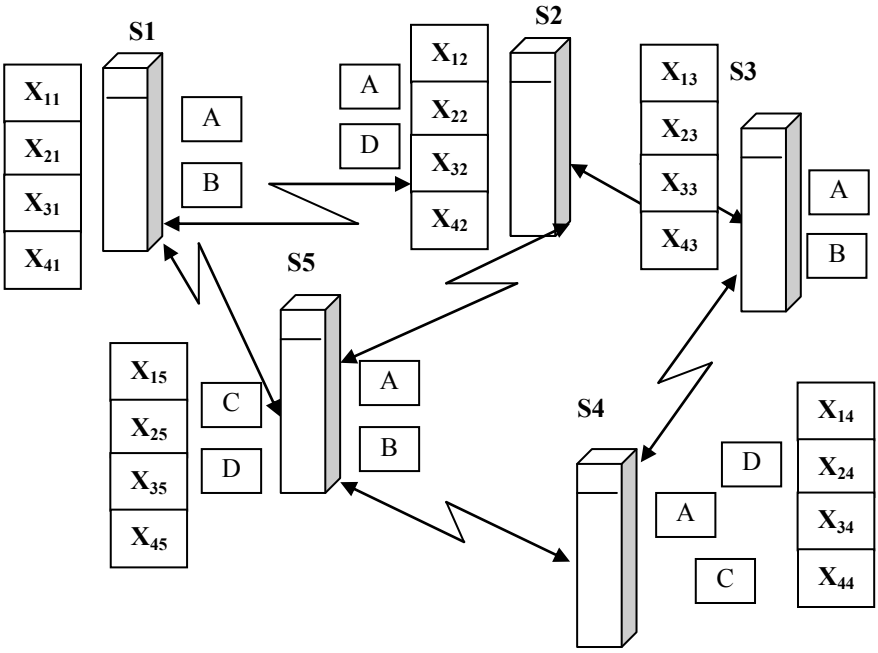

Рис. 2. Схема розміщення компонент бази даних на серверах

Переваги підходу – забезпечений максимальний доступ усіх кінцевих користувачів до даних, які розподілені по вузлах мережі, що забезпечує зручність роботи з даними і максимально високий час відгуку системи на запити. Недоліки – існує проблема «поширення відновлення» [5].

Висновки. Приведено математичні моделі, що дозволяють виконати розподіл даних у корпоративній інформаційній системі на підставі класичних критеріїв розподілу даних таких як мінімізація дублювання даних, максимально можливе наближення даних до місць їхнього використання, мінімізація часу відгуку системи в додатках і відомих підходів до рішення задач оптимізації. Моделі є ідеалізованими і не дозволяють врахувати безліч обмежень, що існують у реальних інформаційних системах, проте вони можуть бути основою для рішення задач організації даних у розподілених інформаційних системах.

ЛІТЕРАТУРА

1. *Саймон А.Р. Стратегические технологии баз данных: менеджмент на 2000 год. – М.: Финансы и статистика, 1999. – 479 с.*
2. *Пономаренко В.С., Павленко Л.А. Організація даних у розподілених інформаційних системах: Навчальний посібник. – Х.: ХДЕУ, 2000. – 104 с.*
3. *Пономаренко В.С., Павленко Л.А. Інструментальні засоби розробки та підтримки баз даних розподілених інформаційних систем: Навчальний посібник. – Х.: ХДЕУ, 2001. – 132 с.*
4. *Павленко Л.А. Корпоративні інформаційні системи: Навчальний посібник. – Х.: ІНЖЕК, 2005. – 260 с.*
5. *Дейт К. Дж. Введение в системы баз данных: 6-е изд. – К.: Диалектика, 1998. – 784 с.: ил.*
6. *Кренке Д. Теория и практика построения баз данных: 9-е изд. – С.-Пб.: Питер, 2005. – 859 с.*
7. *Акоф Р., Сасиени М. Основы исследования операций. – М.: Мир, 1971. – 534 с.*

Надійшла 1.09.2005

Рецензент: доктор технічних наук, професор В.П. Авраменко,
Харківський національний університет радіоелектроніки.