

ВИКОРИСТАННЯ СУЧАСНОЇ ТЕХНОЛОГІЇ РОЗПОДІЛЕНИХ БАЗ ДАНИХ ПРИ ПРОЕКТУВАННІ ІНФОРМАЦІЙНИХ СИСТЕМ

Солдатко Р.В.

*Мелітопольський державний педагогічний університет
імені Богдана Хмельницького, м. Мелітополь*

Сіциліцин Ю.О.

*старший викладач кафедри інформаційних технологій,
Таврійський державний агротехнологічний університет
e-mail: yurarud@mail.ru*

Постановка проблеми. З появою електронних обчислювальних машин (ЕОМ) різко розширилися рамки можливостей людини управляти інформаційними потоками даних. У наш час на передній план виходить правильне і грамотне оволодіння інструментами і методами вирішення завдання накопичення, управління інформацією та швидкого доступу до неї. Виникає необхідність створення єдиного інформаційного простору. Обсяг інформації, що надходить до людини, безперервно зростає. Тому для кожної людини важливо володіти засобами доступу до даних. Для більш успішної роботи підприємства необхідно розглядати інструменти роботи з базою даних з використанням мережі Internet. Під інструментом мається на увазі спеціальне програмне забезпечення. Архітектура програми - клієнт-серверна. Бази даних розташовуються на виділених комп'ютерах іменовані серверами. Це служить джерелом розподіленої архітектури. Логічно зв'язану інформацію представлену у вигляді баз даних з таблицями необхідно децентралізувати. Це досягається шляхом поділу бази даних, або навіть однієї таблиці бази даних, на декілька серверів. У свою чергу сервера можуть бути децентралізовані географічно. Робиться це для того, що б рівномірно розподілити навантаження на таблиці баз даних. Постає проблема правильного розподілення таблиць між серверами.

Аналіз останніх досліджень і публікацій. За останні декілька років розподілені і паралельні СУБД стали реальністю. Вони надають функціональність централізованих СУБД, але у середовищі, де дані розподілені між комп'ютерами, пов'язаними мережею, або між вузлами багатопроцесорної системи. У паралельній системі баз даних правильне розміщення даних має вирішальне значення для забезпечення балансу завантаження. В ідеалі можна повністю уникнути протидії між одночасно виконуваними паралельними операціями, якщо кожна з них буде працювати з незалежною безліччю даних. Незалежні множини даних виходять шляхом декластеризації (горизонтальної фрагментації) відносин на основі функції (хеш-функції або індексації по діапазону) від одного або більше атрибутів і розміщення фрагментів на окремих серверах. Декластеризація може проводитися по одному або декільком атрибутам. В останньому випадку [5] запит точного зіставлення з декількох атрибутів може виконуватися на одному вузлі без взаємодій з іншими вузлами. Вибір між хешуванням і

індексацією по діапазону - це питання організації бази даних. Декластеризація, яка спочатку була запропонована для систем без поділу ресурсів, виявилася корисною і для систем з пам'яттю, що розділяється, оскільки вона сприяє зниженню конфліктності при доступі до пам'яті [6].

Виділення не вирішених раніше частин загальної проблеми. У порівнянні з локальними базами даних обробка розподілених запитів займає окрему не до кінця досліджену нішу в області баз даних. Відомо, що обробка запитів в розподіленій базі даних є більш складною у порівнянні з запитами до локальної бази даних. Складність полягає в тому, що б скоротити час обробки запитів до розподіленої бази даних за допомогою оптимального розміщення вузлів при найбільш оптимальному розміщенні таблиць по різних серверах. У розглянутих працях добре подана інформація про розподілені обчислення у мережі. Але не акцентовано увагу на правильному розподілі таблиць баз даних між вузлами.

Мета статті. Головною метою цієї роботи є виділення основних аспектів, на які слід звертати увагу при проектуванні розподілених баз даних. Скласти математичну модель розміщення таблиць розподілення бази даних по вузлах розподіленої мережі.

Виклад основного матеріалу. Припустимо що розподілена мережа складається з деякої кількості локальних мереж, що утворюють вузли, з'єднаних між собою мережею Internet. Запит, що надходить у базу даних будь-якого вузла, припускає доступ до певної таблиці РБД. Запит до РБД надходить до вхідної черги відповідного вузла мережі. Якщо необхідна для обробки запиту таблиця знаходиться в базі локального вузла, то запит обробляється і видається результат. Якщо таблиця знаходиться на віддаленому вузлі - запит пересилається в цей вузол, обробляється в ньому і результат відправляється у вузол, з якого стався запит.

В описаній схемі обробки запитів в системі впливає, що в процесі обслуговування протягом кожної одиниці часу по мережі пересилається деякий обсяг даних. Загальний обсяг запитів і результатів залежить від розподілу таблиць між вузлами. Чим менше середній обсяг даних, що пересилаються - тим вище буде швидкість обробки запитів.

Припустимо,

n - кількість вузлів мережі;

m - кількість таблиць РБД;

K_j - j -ий вузол мережі;

F_i - i -а таблиця РБД;

L_i - обсяг i -ої таблиці;

b_j - обсяг пам'яті вузла K_j , призначеної для розміщення таблиць;

l_{ij} - інтенсивність запитів до таблиці F_i , ініційованих у вузлі K_j ;

a_i - обсяг запиту до таблиці F_i ;

b_i - обсяг запитуваних даних при виконанні запиту до таблиці F_i ;

Тоді обсяг даних, що надходять на термінал вузла, що містить i -у таблицю, при виконанні запиту до таблиці F_i , дорівнює $\alpha_i + \beta_i$. При цьому якщо x_{ij} ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$) – величини, що визначаються за формулою

$$x_{ij} = \begin{cases} 1, \text{якщо таблиця } F_i \text{ знаходиться у вузлі } K_j; \\ 0, \text{якщо таблиця } F_i \text{ не знаходиться у вузлі } K_j, \end{cases}$$

то обсяг даних, що пересилаються при виконанні запиту до таблиці F_i , ініційованого у вузлі K_j , дорівнює $(\alpha_i + \beta_i)(1 - x_{ij})$.

Оскільки λ_{ij} – інтенсивність запитів до таблиці F_i , ініційованих у вузлі K_j , то вона породжує обсяг даних $\lambda_{ij}(\alpha_i + \beta_i)(1 - x_{ij})$, що потребують пересилання. Тому загальний обсяг даних, які необхідно переслати по каналах зв'язку між вузлами внаслідок функціонування системи в перебігу одиниці часу, визначається формулою

$$S = \sum_{i=1}^m \sum_{j=1}^n \lambda_{ij}(\alpha_i + \beta_i)(1 - x_{ij})$$

якщо підставити

$$\lambda = \sum_{i=1}^m \sum_{j=1}^n \lambda_{ij}$$

то середній обсяг даних, необхідних для пересилання по лініях зв'язку при виконанні запиту в системі дорівнює

$$V = \frac{1}{\lambda} \sum_{i=1}^m \sum_{j=1}^n \lambda_{ij}(\alpha_i + \beta_i)(1 - x_{ij}) \quad (1)$$

Отже, чим менше значення V , тим вище швидкість обробки запитів. Оскільки кожна таблиця F_i ($i = 1, 2, \dots, m$) знаходиться на одному з вузлів мережі, то

$$\sum_{j=1}^n x_{ij} = 1 \quad (i = 1, 2, \dots, m) \quad (2)$$

Крім того, обсяг локальної бази даних кожного вузла K_j ($j = 1, 2, \dots, n$) не повинен перевищувати обсяг пам'яті цього вузла, призначений для розміщення таблиць баз даних. Тому

$$\sum_{i=1}^m L_i x_{ij} \leq b_j \quad (j = 1, 2, \dots, n) \quad (3)$$

Таким чином, завдання оптимального розподілу таблиць по вузлах полягає в тому, щоб визначити значення змінних x_{ij} , де

$$x_{ij} = \{0 \cup 1\} \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, n)$$

які задовольняють умовам (2), (3) і дають мінімум лінійної функції (1). Отримана математична модель є завданням цілочисельного лінійного програмування з булевими змінними.

Запити до однієї і тієї ж таблиці F_i , вступники на сервера різних вузлів, можуть бути різних типів. У цьому випадку обсяг запиту і обсяг запитуваних даних залежить від вузла, на який надійшов запит.

Припустимо, a_{ij} – обсяг запиту до таблиці F_i , ініційованого на сервері вузла K_j ; b_{ij} – обсяг запитуваних даних при виконанні запиту до таблиці F_i , надійшов на сервер вузла K_j . Тоді цільова функція задачі

оптимального розподілу файлів по вузлах обчислювальної мережі може бути представлена у вигляді

$$V = \frac{1}{\lambda} \sum_{i=1}^m \sum_{j=1}^m \lambda_{ij} (\alpha_i + \beta_i) (1 - x_{ij})$$

Запишемо цільову функцію задачі оптимального розподілу таблиць по вузлах у випадку, коли запити з фіксованого вузла до однієї і тієї ж таблиці можуть бути різного типу.

Припустимо,

S - число типів запитів;

λ_{kij} - інтенсивність запитів k -го типу до таблиці F_i , ініційованих у вузлі K_j ;

α_{kij} - обсяг запиту k -го типу до таблиці F_i , ініційованого на сервері вузла K_j ;

β_{kij} - обсяг запитуваних даних при виконанні запиту k -го типу до таблиці F_i , надійшов на сервер вузла K_j ;

Тоді обсяг даних, що пересилаються при виконанні запиту k -го типу до таблиці F_i , ініційованого у вузлі K_j , дорівнює $[(\alpha)_{kij} + \beta_{kij}](1 - x_{ij})$. Оскільки інтенсивність λ_{kij} породжує обсяг даних $\lambda_{kij}[(\alpha)_{kij} + \beta_{kij}](1 - x_{ij})$, потребують пересилання, то загальний обсяг даних, які необхідно переслати по каналах зв'язку між вузлами унаслідок функціонування розподіленої системи протягом одиниці часу, визначається формулою

$$S = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^s \lambda_{kij} (\alpha_{kij} + \beta_{kij}) (1 - x_{ij})$$

Якщо підставити

$$\lambda = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^s \lambda_{kij}$$

то цільова функція задачі оптимального розподілу файлів по вузлах обчислювальної мережі прийме вигляд

$$V = \frac{1}{\lambda} \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^s \lambda_{kij} (\alpha_{kij} + \beta_{kij}) (1 - x_{ij})$$

Всі повідомлення, що надходять у вхідні черги вузлів, розділимо на два типи: *тип 1* - повідомлення, складові запити, для обробки яких необхідні таблиці не містяться в базі даних вузла; *тип 2* - повідомлення, що складають запити, для обслуговування яких потрібні таблиці містяться в базі даних відповідного вузла. При цьому будемо вважати, що запит *типу 1*, прибувши-ший для свого обслуговування у віддалений вузол, перетворюється на запит *типу 2*.

Введемо величини:

W_{sk} - середній час обробки повідомлення k -го ($k = 1, 2$) типу у вузлі K_s ;

U_{sk} - інтенсивність надходження повідомлень k -го ($k = 1, 2$) типу до вхідної черги вузла K_s .

Тоді, аналогічно замість обмеження (3) в математичній моделі можна використовувати наступне обмеження

$$U_{s1}W_{s1} + U_{s2}W_{s2} < 1,$$

де

$$U_{s1} = 2 \sum_{i=1}^m \lambda_{is}(1 - x_{is}) \quad (s = 1, 2, \dots, n);$$

$$U_{s2} = \sum_{i=1}^m \lambda_i x_{is} \quad (s = 1, 2, \dots, n);$$

$$\lambda_i = \sum_{j=1}^n \lambda_{ij}$$

Розглянемо побудова моделі оптимального розподілу таблиць по вузлах обчислювальної мережі у випадку, коли в якості критерію оптимальності приймається загальна вартість трафіку, породженого функціонуванням обчислювальної системи протягом одиниці часу.

Припустимо,

n - число вузлів мережі;

m - число незалежних файлів РБД;

K_j - j -ий вузол мережі;

F_i - i -а таблиця РБД;

L_i - обсяг i -ої таблиці;

b_j - обсяг пам'яті вузла K_j , призначеної для розміщення таблиці;

λ_{ij} - інтенсивність запитів до таблиці F_i , ініційованих у вузлі K_j ;

α_{ij} - обсяг запиту до таблиці F_i , ініційованого на сервері вузла K_j ;

β_{ij} - обсяг запитуваних даних при виконанні запиту до таблиці F_i , що надійшов на сервер вузла K_j ;

r_{sj} - вартість передачі одиниці інформації з вузла K_s у вузол K_j .
($r_{ss} = 0, s = 1, 2, \dots, n$);

Якщо x_{ij} ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$) - величини, що визначаються за формулою

$$x_{ij} = \begin{cases} 1, \text{ якщо таблиця } F_i \text{ знаходиться у вузлі } K_j; \\ 0, \text{ якщо таблиця } F_i \text{ не знаходиться у вузлі } K_j, \end{cases}$$

тоді обсяг даних, що пересилаються при виконанні запиту до таблиці F_i , ініційованого у вузлі K_j , дорівнює $(\alpha_{kij} + \beta_{kij})(1 - x_{ij})$. Оскільки інтенсивність λ_{ij} породжує обсяг даних $\lambda_{ij}(\alpha_{ij} + \beta_{ij})(1 - x_{ij})$, потребують пересиланні, то вартість пересилки цих даних дорівнює

$$C_{ij} = \sum_{s=1}^n \lambda_{ij} (\alpha_{ij} + \beta_{ij}) r_{sj} (1 - x_{ij}) x_{is},$$

а загальна вартість даних, які слід переслати по лініях зв'язку між вузлами внаслідок функціонування системи протягом одиниці часу, визначається формулою

$$C = \sum_{i=1}^m \sum_{j=1}^n \sum_{s=1}^n \lambda_{ij} (\alpha_{ij} + \beta_{ij}) r_{sj} (1 - x_{ij}) x_{is}$$

Таким чином, завдання полягає в тому, щоб знайти мінімум функції вартості C при обмеженнях (2), (3).

Висновки і пропозиції. Проведено огляд переваг використання розподілених баз даних при проектуванні інформаційних систем для підприємства. У даній статті розроблено математичну модель загальної вартості пересилання даних у мережі між вузлами. За допомогою цієї функції можна розподілити навантаження на сервера баз даних, тим самим зменшити очікування користувача. Слід акцентувати увагу на правильному розміщенні вузлів та таблиць бази даних. Вузли, до яких йде більш часте звернення потрібно виділяти на більш потужні сервера. Таблиці, між якими ведеться часте звернення, рекомендовано розміщати на одному вузлі. Чим менше буде вартість пересилання даних – тим менш часу потребується на виконання.

Література

1. Глоба Л.С. Розробка інформаційних ресурсів та систем / Л.С. Глоба // Київ, 2013. – 433 с.
2. Избачков Ю.С., Петров В.Н. Информационные системы / Ю.С. Избачков - СПб.: 2006. – 656 с.
3. Музылева И.В. Основы цифровой техники / И.В. Музылева // ИНТУИТ, 2011. - 354 с.
4. Таненбаум Э., Ван Стеен М. Распределенные системы. Принципы и парадигмы / Э. Таненбаум - СПб.: Питер, 2003. – 877 с.

Анотація. У статті розкривається питання актуалізації використання сучасної технології розподілених баз даних при проектуванні інформаційних систем для роботи підприємств. Запропоновано математичну модель розподілу навантаження. Виділяються основні складові параметри пересилання даних які впливають на час доставки даних по мережі. Поставлено проблема про те, як фізично розподіляти базу даних по серверах. Зроблено висновок про те, як краще *розподілити навантаження на сервера баз даних*.

Ключові слова: інформаційна система, бази даних, проектування, ЕОМ, запит, автоматизація, математична модель.

Аннотация. В статье раскрывается вопрос актуализации использования современной технологии распределенных баз данных при проектировании информационных систем для работы предприятий. Предложена математическая модель распределения нагрузки. Выделяются основные составляющие параметры пересылки данных, которые влияют на время доставки данных по сети. Поставлена проблема о том, как физически распределять базу данных по серверам. Сделан вывод о том, как лучше распределить нагрузку на сервера баз данных.

Ключевые слова: информационная система, базы данных, проектирование, ЭВМ, запрос, автоматизация, математическая модель.

Abstrac. The article deals with the issue of mainstreaming the use of modern technology of distributed databases in the design of information systems for businesses. A mathematical model of load distribution. There are the basic components of data transmission parameters that affect the delivery of data over the network. The problem of how to physically distribute the database on servers. The conclusion about how best to distribute the load on the database server.

Keywords: information system, database design, computer request, automation, mathematical model.