МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА ОБЧИСЛЮВАЛЬНІ МЕТОДИ

УДК 004.658 + 652.3

М. Т. Фісун, М. Л. Дворецький, С. В. Дворецька

ПОБУДОВА МОДЕЛЕЙ ДЛЯ ОПТИМІЗАЦІЇ СТРУКТУРИ БАЗИ ДАНИХ ВУЗЛА У КОРПОРАТИВНИХ ІНФОРМАЦІЙНИХ СИСТЕМАХ

Чорноморський національний університет ім. П. Могили, Миколаїв

Анотація. Обгрунтовано актуальність роботи, яка визначається тенденцією переходу від локальних до розподілених баз даних та недоліками використання "універсальних" облікових систем. Аргументовано важливість питання правильного розподілення даних між вузлами розподілених та територіально розосереджених комп'ютерних систем.

При постановці задачі та визначенні мети дослідження увага звертається на необхідність визначення та побудови математичної моделі критеріїв оптимальності структури віддаленого вузла розподіленої БД. Враховується статистика користувацької активності, що представляється у вигляді багатовимірної моделі даних та для кожного елементу вимірів визначається маркер представленості даних на віддаленому вузлі. Виконавши відповідні операції консолідації для рівня атрибутів кортежів відношення, відповідно до запропонованої функції агрегації, отримується значення маркеру представленості кожного з них.

На наступному етапі сформульовано критерії ефективності моделі: незалежність від центрального вузла БД, розмір локальної БД та показник рівня необхідності синхронізації даних. Побудовано математичні залежності між їх значеннями та значенням граничного рівня представленості для даних віддаленого вузла. Це дає змогу, розв'язавши задачу багатокритеріального аналізу, визначити рівень представленості даних, для якого значення критеріїв ефективності відповідатимуть координатам найоптимальнішої точки простору рішень.

Ключові слова: система керування розподіленими БД, багатокритеріальна оптимізація, реляційна модель даних, SQL-запит.

Аннотация. Во вступительной части обоснована актуальность работы, которая определяется тенденцией перехода от локальных к распределенным базам данных и недостатками использования "универсальных" учетных систем. Аргументировано важность вопроса правильного распределения данных между узлами распределенных и территориально рассредоточенных КС.

При постановке задачи и определении цели исследования внимание обращается на необходимость определения и построения математической модели критериев оптимальности структуры удаленного узла распределенной БД. Учитывается статистика пользовательской активности, которая представляется в виде многомерной модели данных и для каждого элемента измерений определяется маркер представленности данных на удаленном узле. Выполнив операции консолидации для уровня атрибутов кортежа отношения, согласно предложенной автором функции агрегации, получается значение уровня необходимости представления каждого из них.

На следующем этапе сформулированы критерии эффективности модели (независимость от центрального узла БД, размер локальной БД, и показатель уровня необходимости синхронизации данных) и построены математические зависимости между их значениями и значением предельного уровня представленности для данных удаленного узла. Это позволяет, решив задачу многокритериального анализа, определить уровень представленности данных, для которого значения критериев эффективности будут отвечать координатам оптимальной точки пространства решений.

Ключевые слова: распределенная СУБД, многокритериальная оптимизация, реляционная модель данных, SQL-запрос.

Abstract. The introductory part of the research justifies the relevance of the work. Because of the tendency to move from local to distributed databases, and the disadvantages of using "universal" accounting systems many companies use the set of special solutions instead. According to this, it is very important to distribute data of corporate database between its nodes correctly. The key factor that affects the reliability and availability of the database is the link localization. If the database is distributed optimally, the data hosted in this node is called exclusively by its user.

The research aims to determine and build a mathematical model of the optimality criteria for the structure of a remote node of the distributed database. The statistics of user SQL-queries activity is taken into account. Statistics results is presented in the form of a multidimensional data model. It this model each dimension element is marked with a marker of data presentation on the remote node. Next, the statistics data is consolidated according to the aggregation function proposed by the author. As a result, the necessity level value of data cell representing is obtained.

At the next stage, criteria the model effectiveness are formulated (independence from the central node of the database, the size of the local database, and an indicator of the level of need for data synchronization). Then mathematical dependencies between criteria values and the value of data representation level on the remote node are built. This allows, having solved multicriteria analysis problem, to determine the level of data representation for which the values of the criteria will correspond to the coordinates of the optimal point in the solution space.

Keywords: distributed database management system, multi-criteria optimization, relational data model, SQL-query.

DOI: https://doi.org/10.31649/1999-9941-2020-48-2-52-60.

Вступ

У розвитку сучасних інформаційних систем вже давно намітилася тенденція переходу від локальних до створення розподілених баз даних (РБД). На ринку програмних засобів існує чимало систем керування базами даних (СКБД), що дають змогу розміщувати, підтримувати та обробляти дані на різних вузлах комп'ютерних інформаційних систем (КІС). В подальшому такі СКБД будемо називати розподіленими і позначати СКРБД. Основною задачею СКРБД є забезпечення управління доступом до даних багатьох користувачів і забезпечення цілісності та узгодженості даних в умовах використання мережі користувацьких пристроїв [1]. Так, у межах одного підприємства або компанії існує необхідність автоматизації різних типів обліку [2, 3]. У якості прикладу можна навести: складський облік, бухгалтерський облік, облік кадрів, розробка інформаційних порталів, системи відеоспостереження,

контроль прав доступу до даних та ін. Спроба автоматизації всіх видів обліку привело до виникнення так званих «універсальних» або «комплексних» облікових систем [3], що створюють єдине облікове середовище організації та забезпечують доступ до всіх необхідних даних моніторингу, контролю, аналізу й оцінки ефективності роботи організації, а також для підтримки прийняття управлінських рішень. Даний підхід має цілий ряд недоліків [2,4], позбавитись яких дозволяє використання окремих спеціалізованих рішень [3,5]. Але такий шлях веде до появи різних СКБД, що потребують їх синхронізації [6]. Задачею систем керування розподіленими БД є забезпечення управління доступом до даних багатьох користувачів, цілісності й узгодженості даних в умовах використання комп'ютерної мережі (КМ). Тобто,поруч з основною функцією СКБД: уведення, зберігання, пошук, обробка, сумісне користування даними та централізоване керування – специфічною важливою функцією є забезпечення спільної роботи багатьох користувачів з розподіленою інформацією [7,8].

Актуальність

Ключовим фактором, що впливає на надійність і доступність до бази даних, є так звана локалізація посилань [5]. Якщо база даних розподілена так, що дані, які розміщені в якомусь вузлі, викликаються винятково його користувачем, то це свідчить про високий рівень локалізації посилань. Якщо подібне розподілення даних здійснити неможливо і для виконання запитів користувача потрібно звертатись, переважно, за інформацією і до інших вузлів, то це свідчить про невисокий рівень локалізації посилань.

Комбінована стратегія розподілу даних є найбільш виправданою із точки зору можливості поєднання переваг стратегій з/без дублювання. Але при її використанні, окрім задачі синхронізації дубльованої інформації, актуальною постає задача проектування структури БД з точки зору приналежності даних до того чи іншого вузла мережі. Крім того, продуктивність системи напряму буде залежати від прийняття рішення щодо необхідності часткового або повного дублювання даних. Деякі відношення реляційної БД можуть бути продубльовано повністю, а деякі – після застосування операцій проекції та вибірки. Тобто, для оптимізованого представлення даних на віддаленому вузлі необхідно використати процедури вертикальної та горизонтальної фрагментації даних.

Отже, досить важливим є питання раціонального розподілення даних між вузлами розподілених та територіально розосереджених КІС. Тому задача оптимізації структури БД територіально віддаленого вузла у корпоративних інформаційних системах представляється актуальною.

Мета

Для ефективного розв'язання задачі багатокритеріальної оптимізації структури БД віддаленого вузла розподіленої та територіально розосередженої КІС необхідно, передусім, побудувати математичну модель оптимізації, що є *метою* даної статті. При цьому дослідження стосується тільки БД, побудованих на реляційній моделі даних.

Реляційна модель даних основана на простому й у той же час потужному математичному апараті, що спирається, головним чином, на теорію множин і математичну логіку [9,10]. При побудові математичної моделі розподіленої БД, а також формалізації роботи розподіленої транзакції, механізмів синхронного та асинхронного оновлення даних [11], extractingtransformationloading (ETL) процесів вважається за доцільне використання базових понять теорії множин.

Розроблювана модель має враховувати статистику користувацької активності звернення до локальних та віддалених даних, на основі якої, використовуючи фільтрацію за обраними для аналізу вимірами, можуть бути отримані відповідні зрізи або підмножини даних [12]. Для елементів вимірів уведемо характеристика «маркер представленості даних», що визначає ступінь їх необхідності на вузлі розподіленої корпоративної інформаційної системи (РКІС). Від агрегованого значення даного маркеру РКІС на відповідній віддаленому вузлу підмножині БД будуть залежати значення критеріїв ефективності моделі, серед яких незалежність від центрального вузла БД, розмір локальної БД та показник рівня необхідності синхронізації даних [13]. Отже, метою дослідження є математичне представлення залежності значень вищенаведених критеріїв оптимальності від значення маркеру представленості даних.

Розв'язання задач

Серед загальновідомих операцій реляційної алгебри [9], зважаючи на горизонтальну та вертикальну фрагментацію даних на вузлі РКІС, тут розглядаються операції «проекція» і «вибірка». Нехай tup – кортеж відношення R, tup[P] – частина цього кортежу, що містить тільки значення атрибутів, які входять до підмножини P схеми відношення R_{shema} ($P \subset R_{shema}$). Тоді проекцією R на P будуть відношення, що складаються з кортежів усіх значень з множини P, що існують у відношенні R, тобто $R[P] = \{tup[P] \mid tup \in R_{data}\}$. Схема результуючого набору рядків може бути визначена наступною множиною атрибутів: $R[P]_{shema} = \{A_1 \ldots, A_m\}$, де $A_i \in R_{shema}$.

Вибірка виконує відображення кортежів, результатом якого є відношення, що містить підмножину всіх унікальних кортежів відношення R, для яких виконується певна логічна умова: $R[S] = \{ tup \mid tup \in R_{data} \wedge F(tup, S) = істина \}$, де S – деяка логічна умова, а F(tup, S) – функція, що відображає її

виконання для відповідного кортежу. Схема результуючого набору рядків буде збігатись зі схемою базового відношення, тобто $R[S]_{shema} = R_{shema}$.

У межах SQL – запиту на вибірку даних може приймати участь деяка множина відношень, кожне з яких є результатом послідовного виконання операцій вибірки та проекції до базового відношення (таблиці БД). R'' = R'[P], де R' = R[S], тобто

$$R'' = \{tup[P] \mid tup[P] \in R[P]_{data} \wedge F(tup, S) = істина\} \quad (1)$$

Розглядаючи множину запитів до БД, результуюча підмножина R''_{union} базового відношення R може бути визначена як об'єднання підмножин R' всіх запитів, що надійшли до БД з віддаленого вузла

$R''_{union} = \bigcup_{i=1}^{n} R_i''$, або

$R''_{union} = \{tup[P_{union}] \mid tup[P_{union}] \in R[P_{union}]_{data} \wedge F(tup, S_{union}) = істина\}$, де

$tup[P_{union}] = \bigcup_{i=1}^{n} tup[P_i]$, а $S_{union} = \bigvee_{i=1}^{n} S_i$

Враховуючи, що деякі дані, необхідні на вузлі РБД, для запобігання необхідності подальшої реплікації, можуть бути представлені лише на центральному вузлі БД та приймати участь у роботі за рахунок використання розподілених запитів, результуюче відношення R^{remote} буде лише підмножиною R''_{union}. Зважаючи на той факт, що для представлення даних на віддаленому вузлі територіально розосередженої КІС необхідно використати елементи як вертикальної, так і горизонтальної фрагментації даних (як проекцію, так і вибірку), підмножина базового відношення R, що буде описувати відношення віддаленого вузла, може бути представлена наступним чином:

$$R^{remote}_{shema} = \{A \mid A \in R_{shema}, R_{primary} \subset R^{remote}_{shema}, A \in R_{primary} \vee F_a(Node, A) = істина\} \quad (2)$$

Для прийняття рішення про представленість певного атрибуту відношення на тому чи іншому вузлі використовуємо функцію F_a(Node, A). Зауважимо, що множина атрибутів первинного ключа відношення у будь-якому випадку має бути представлена на віддаленому вузлі.

Множина кортежів, у свою чергу, визначатиметься за формулою:

$$R^{remote}_{data} = \{ tup \mid tup \in R_{data}, tup_{primary} \in R^{remote\text{-}dep}_{data} \vee F_{tup}(Node, tup) = істина\} \quad (3)$$

Як бачимо, кортеж має бути представлений у випадку входження його первинного ключа до множини даних відношень, залежних від поточного. В іншому випадку необхідність наявності даних вирішується за допомогою оціночної функції F_{tup}(Node, tup).

Модель представлення користувацьких запитів має підтримувати можливість їх подальшої класифікації згідно приналежності до того чи іншого автоматизованого робочого місця, географічного розташування, ролі користувача та інших критеріїв, що можливо додати до моделі під час її використання згідно з особливостями тієї чи іншої предметної області. Тобто, користувацький запит визначається, як

$$Q = <АРМ, Користувач, Застосунок, R_{set''}, Q_{set}^{inner}>, \quad (4)$$

де АРМ = <Тип, Розташування> – автоматизоване робоче місце, що характеризується типом та розташуванням;Користувач = <Роль, Ім'я> – користувач корпоративної ІС із ім'ям та роллю; $R_{set}'' = \{ R'' \mid \{tup[P] \mid tup[P] \in R[P]_{data} \wedge F(tup, S) = істина\} \}$ – множина результуючих відношень, отриманих із базових відношень (таблиць) БД відповідними запитами; Q_{set}^{inner} – множина вкладених запитів основного запиту Q. Зауважимо, що тип та розташування АРМ, ролі користувачів та додаткові властивості задаються за допомогою відповідних класифікаторів.

При плануванні структури БД віддаленого вузла територіально розосередженої КІС та, відповідно, розрахунку значень функцій F_a(Node, A) та F_{tup}(Node, tup), що визначають необхідність представлення даних локально, прийматиме участь декілька факторів – доступність та швидкість отримання даних, незалежність від центрального вузла БД, розмір БД, ступінь достовірності даних, необхідність у подальшій синхронізації.

На першому кроці моделювання розпочинається з представлення у віддаленому вузлі повної копії БД центрального вузла. При цьому доступність та незалежність від центрального вузла БД має максимальний рівень, швидкість отримання даних у порівнянні із центральним вузлом, як правило, нижча за рахунок менш потужних обчислювальних ресурсів, але може бути збільшена за рахунок виконання операцій вибірки та проекції до відношень та зменшення кількості та обсягів блокувань даних. Розмір БД є великим, а отже даний критерій є неоптимальним. Також всі дані потребують синхронізації з центральним вузлом, що є досить ресурсоємкою операцією.

Другий крок полягає у виключенні з віддаленого вузла всіх надлишкових даних. Для розв'язання даної задачі на базі реляційної моделі даних обліку користувацьких запитів (4) створено багатовимірну модель [14], де у якості базового набору вимірів запропоновано:

$$D = <ДатаЧас, ТипАРМ, РозташуванняАРМ, Роль, Застосунок, R, A, tup> \quad (5)$$

Множина мір складається з пари значень «Кількість» та «Час Виконання»

$$M = <Кількість, Час Виконання> \quad (6)$$

Домени вимірів «Тип АРМ», «Роль Користувача» та «Застосунок» визначаються за допомогою відповідних класифікаторів. При виконанні аналізу щодо представленості даних для елементів вимірів вводиться характеристика маркеру представленості даних, що відображає рівень необхідності представлення даних на вузлі РКІС. Для кожного елементу значення маркеру приймається одне із множини лінгвістичних значень {«необхідно», «бажано», «не потрібно»}, що визначає ступінь необхідності представлення даних того чи іншого типу АРМ, ролі користувача або застосунку. Для виміру «Розташування АРМ» маркування виконується автоматично значенням «необхідно» для відповідного віддаленого вузла, та «не потрібно» для всіх інших.

При визначенні значення маркеру представленості для рядку таблиці фактів [14] використано функцію максимуму, що відображає принцип поглинання. Визначення значення маркеру при виконанні консолідації рядків таблиці фактів по значеннях<R, A, tup>, тобто для комірки таблиці, може бути виконано декількома шляхами. Так, за песимістичними сценарієм використовується функція агрегації «максимум» (у нашому випадку максимальне значення має маркер «необхідно», а відповідно мінімальне «не потрібно»). Даний підхід резервує дані, якщо вони необхідні хоча б в одному місці. За оптимістичним сценарієм, навпаки, дані маркуються як «непотрібно», якщо вони є такими хоча б по одному з вимірів. Зрозуміло, що перший сценарій гарантовано забезпечує віддалений вузол необхідними даними, але веде до потенційної збитковості даних та розширює область даних, що потребуватимуть подальшої синхронізації. Другий у свою чергу веде до мінімізації даних на віддаленому вузлі, але збільшує потребу у розподілених запитах, що у свою чергу веде до зменшення рівня доступності та швидкості отримання даних, а також рівня незалежності від центрального вузла БД.

Застосування методу ковзаючого середнього частково вирішує це питання, але залишається невирішеним питання питомої ваги впливу маркеру кожного з вимірів. Крім того, слід врахувати, що для деяких підмножин вимірів має спрацьовувати саме песимістичний сценарій (дані потрібні, не зважаючи ні на що), а для деяких оптимістичний (дані не мають дублюватися в жодному разі). При чому, перший, як правило, поглинатиме останній.

Отже, маємо модель, у якій кожен атрибут виміру має значення, маркер та ваговий коефіцієнт A_{dim} = ={Val, Mrk, vol}, де Mrk = {«обов'язково», «необхідно», «бажано», «не потрібно», «заборонено»}, а vol – ваговий коефіцієнт (ігнорується для значень маркера «обов'язково» та «заборонено»).

Виконавши переведення нечислової лінгвістичної змінної маркерів у числове значення («обов'язково» – «2», «необхідно» – «1», «бажано» – «0», «не потрібно» – «-1», «заборонено» – «-2»), визначимо для них функцію агрегації:

$$Aggregate_{i=1}^{n} Mrk_i = \begin{cases} 2, \text{ якщо } \exists\, Mrk_i = 2 \\ -2, \text{якщо } \exists\, Mrk_i = -2 \ \wedge \nexists\, Mrk_i = 2 \\ \sum_{i=1}^{n} (Mrk_i * \frac{Vol_i}{\sum_{i=1}^{n} Vol_i}) \end{cases} \quad (7)$$

При прийнятті рішення щодо представленості даних на віддаленому вузлі виконуємо консолідацію рядків таблиці фактів за множиною <R, A, tup>та розраховуємо значення маркеру для кожного її елементу за формулою (7), після чого приймаємо рішення про представленість за наступним критерієм.

$$Предст(Node,R,A,tup) = (Aggregate(R, A, tup)_{i=1}^{n} Mrk_i > koef_{\text{предст}}^{node}), \quad (8)$$

де $koef_{\text{предст}}^{node}$ – пороговий коефіцієнт представленості даних у певному вузлі Node, область значень якого [-1, 1].

При встановленні коефіцієнту представленості даних (8) значення одиниці ($koef_{\text{предст}}^{node} = 1$) розміщуємо всі дані, необхідні для роботи віддаленого вузла РБД у локальній базі даних. Звідси, як і у випадку повного дублювання даних, маємо той же максимальний рівень доступності даних. Також, у порівнянні із повним дублюванням, не змінюється рівень достовірності та незалежності від центрального

вузла БД. Але швидкість отримання даних збільшується за рахунок зменшення їх обсягу (передусім на операціях вибірки). Також маємо менший розмір БД та зменшення обсягів даних, що потенційно можуть потребувати подальшої синхронізації.

Третій крок полягає у повній відмові від локальної БД та розміщенні всіх даних та центральному вузлі (або, в окремих випадках, в інших вузлах) РКІС. У цьому випадку маємо максимізацію рівня оптимальності за показниками щодо необхідності подальшої синхронізації даних, оскільки дублювання даних відсутнє. Рівень достовірності також є максимальним, і розмір локальної БД має мінімальне значення (локальна БД відсутня). Але,у той же час, максимально знижується доступність даних та швидкість їх отримання, а робота КІС є максимально залежною від доступності центрального вузла.

Відмітимо, що на третьому кроці значення частини показників у порівнянні із другим кроком було покращено, але у той же час значення іншої частини показників погіршилося. Логічно припустити, що оптимальних значень показники якості структури вузла РКІС набувають на проміжку між 2-м та 3-м кроками. Для можливості виконання аналізу та знаходження оптимального розподілу даних між віддаленим та центральним вузлами необхідно, передусім, формалізувати перераховані вище показники якості структури БД.

Показники незалежності від центрального вузла БД, та, відповідно, доступності і швидкість їх отримання напряму залежать від представленості даних користувацьких запитів на вузлі РКІС. Використавши модель запиту користувача (4) та результуючого відношення віддаленого вузла (1, 2), можемо визначити функцію доступності даних запиту:

$$F_{доступність}(Node, Q) = \begin{cases} 1, \text{якщо } \forall R'' \exists R''_{shema}, R'' \in R^{remote}_{shema} \wedge \\ \quad \forall Q^{inner} F_{\text{доступність}}(Q^{inner}) = 1 \\ 0, \text{ якщо } \exists R'' \nexists R^{remote}_{shema}, R'' \in R^{remote}_{shema} \vee \\ \quad \exists Q^{inner} F_{\text{доступність}}(Q^{inner}) = 0 \end{cases} \quad (9)$$

Загальне значення рівня доступності даних та незалежності від центрального вузла БД визначається, як середнє значення рівня доступності (9) деякої підмножини користувацьких запитів.

$$F_{доступність} = \frac{\sum_{i=1}^{n} F_{\text{доступність}}(Q_n)}{n}, \text{ де } Q_n \in Q_{node} \quad (10)$$

Множина користувацьких запитів Q_{node} є підмножиною всіх користувацьких запитів Q_{all}($Q_{node} \subset Q_{all}$), для кожного елементу якої функція приналежності до віддаленого вузла дорівнює одиниці.

$$Q_{node} = \{Q \mid F_{\text{приналежності}}(Node, Q) = 1\},$$

де

$$F_{\text{приналежності}}(Node, Q) = \begin{cases} 1, \text{якщо } (\exists R'' \in R''_{set} \\ \rightarrow Aggregate(R'')_{i=1}^{n} Mrk_i > -1) \vee \\ (\exists Q^{inner} \in Q^{inner}_{set} \\ \rightarrow F_{\text{приналежності}}(\text{Node}, Q^{inner}) = 1) \\ \\ 0, \text{якщо } (\forall R'' \in R''_{set} \\ \rightarrow Aggregate(R'')_{i=1}^{n} Mrk_i \leq -1) \wedge \\ (\forall Q^{inner} \in Q^{inner}_{set} \\ \rightarrow F_{\text{приналежності}}(\text{Node}, Q^{inner}) = 0) \end{cases}$$

Слід зазначити, що до складу елементу Q множини Q_{all} окрім (4) також має бути включений порядковий номер. Дана операція виконується зважаючи на те, що деякі запити надходять від користувачів більше одного разу, і можуть бути згорнуті до одного рядку. Але у нашому випадку, дана особливість має бути врахована, та вага такого запиту при розрахунку загального значення критерію доступності даних віддаленого вузла територіально розосередженої КІС має бути збільшена у порівнянні з іншими запитами відповідно до кількості надходжень,що і досягається за рахунок їх дублювання у множині Q_{all}.

Наступним розглянемо критерій розміру локальної БД. Даний критерій впливає як на продуктивність запитів до локальної БД, так і на потужність обчислювальних ресурсів, необхідних для виконання операцій адміністрування БД та КІС (резервне копіювання, забезпечення відмовостійкості та ін.). База даних під управлінням реляційної СКБД (у тому числі розподіленої) представлена на дисковому просторі у вигляді файлу або групи файлів [7,8]. У той же час будь-яка сучасна реляційна СКБД має у своєму розпорядженні механізми отримання інформації про те, скільки місця на диску займає кожне окреме відношення. В рамках виконання роботи експериментальним шляхом виявлено, що у переважній більшості випадків сумарне значення розміру відношень співпадає або майже співпадає із сумарним значенням розміру файлів БД,

$$\sum_{i=1}^{n} SizeR_i^{DBMS} - \sum_{j=1}^{m} SizeFile_j = \Delta \text{ , де} \frac{\Delta}{\sum_{i=1}^{n} SizeR_i^{DBMS}} \leq 0.005 \qquad (11)$$

Звідки можна зробити висновок про можливість їх використання при розрахунку значення критерію розміру локальної БД віддаленого вузла територіально розосередженої КІС.

Але інформація про розмір R не дає можливості визначити розмір R", що є результатом послідовності операцій вибірки та проекції, і входить до множини R^{remote}. З іншого боку, кожна СКБД надає інформацію про обсяг дискового простору, необхідного для збереження значення атрибуту, визначеного на певному домені [7,8]. Розмір кортежу може бути визначено, як

$$Size_R = SizeR0_i^{DBMS} + p\ x \sum_{i=1}^{n} Size(Type_i), \qquad (12)$$

де $A_i \in D_i \in Type_i$, p – кардинальне число або потужність відношення, а $SizeR0_i^{DBMS}$ – розмір і-го відношення за відсутності даних.

Однак, отримані за допомогою (12) значення не можуть бути використані в розрахунках у чистому вигляді, оскільки $Size_R$ практично ніколи не збігається з $SizeR^{dbms}$, а в деяких випадках відрізняється у декілька разів. Це може бути пов'язано як із наявністю додаткових структур даних (індексів), пов'язаних із відношенням, так і з особливостями побудови кластерних індексів, використання фактору заповнення та інших властивостей представлення даних на диску тих чи інших СКБД.

Одночасне використання (11) та (12) дає можливість подолати вищенаведені суперечності. Так, на першому кроці для кожного відношення визначаємо поправочний коефіцієнт:

$$Koef_{sizeR} = \frac{SizeR_i^{DBMS} - SizeR0_i^{DBMS}}{p \times \sum_{i=1}^{n} Size(Type_i)}.$$

Далі, при визначенні розміру відношення R" (що є підмножиною R) використовуємо наступну формулу

$$Size_{R''} = Koef_{sizeR} \times p' \times \sum_{i=1}^{n'} Size(Type_i), \qquad (13)$$

де p – кардинальне число R", n' – кількість елементів множини $R^{remote}{}_{shema}$ (кількість атрибутів), і кожен атрибут $A_i \in D_i \in Type_i$.

Сума розмірів (13) всіх підмножин R" дасть нам розмір БД віддаленого вузла. Але для кожного окремого випадку предметної області даний розмір буде приймати різні значення, а отже його абсолютне значення не має цінності. Тому було прийнято рішення остаточне значення критерію розміру локальної БД представити у вигляді відношення до загального розміру БД центрального вузла КІС.

$$F_{size} = \sum_{i=1}^{n} \frac{Size_{R'',i}}{SizeR_i^{DBMS}}. \qquad (14)$$

Останнім із зазначених вище критеріїв, є необхідність у подальшій синхронізації. Спочатку визначаємо підмножину даних віддаленого вузла, для якої виконуються операції зміни даних. Для цього визначимо модель SQL-запиту модифікації даних відношення Q^{modif} = <Виміри, R''^{modif}, type>, де R''^{modif} – підмножина на відношенні R, що зазнає змін внаслідок виконання операцій модифікації даних, type = {insert, update, delete} – тип операції. R''^{modif} визначаємо, як

$$R''^{modif} = \{tup[P^{modif}] \mid tup[P^{modif}] \in R[P^{modif}]_{data} \wedge F(tup, S) = \text{істина}\}, \qquad (15)$$

де S – деяка логічна умова, F(tup, S) – функція, що відображає її виконання для відповідного кортежу, а P^{modif} – множина атрибутів, що зазнає модифікації.

Розглядаючи множину запитів до БД, результуюча підмножина R''^{modif}_{node} базового відношення R може бути визначена, як об'єднання підмножин R'' всіх запитів (15), що надійшли до БД з віддаленого вузла$R''^{modif}_{node} = \bigcup_{i=1}^{n} R''_i$, або

$$R''^{modif}_{node} = \{tup[P^{modif}_{node}] \mid tup[P^{modif}_{node}] \in R[P^{modif}_{node}]_{data} \wedge F(tup, S_{node}) = істина\}. \quad (16)$$

Аналогічним чином визначаємо множину R''^{modif}_{main} R"$^{modif}_{main}$, що зазнаватиме модифікацій на центральному вузлі або інших вузлах із подальшою синхронізацією із центральним. Перетин множин R''^{modif}_{node} та R''^{modif}_{main} буде визначати підмножину базового відношення, на якій можуть виникати конфлікти зміни та значення атрибутів якої потребують застосування більш ресурсоємних алгоритмів синхронізації [6, 11, 13].

$$R'''^{modif}_{node} = R''^{modif}_{node} \cap R''^{modif}_{main}. \quad (17)$$

Ґрунтуючись на (16, 17) додаємо до багатовимірної моделі (5, 6) вимір SyncroFlg = {true, false}, що визначатиметься на кортежі<R, A, tup>. Далі, на підставі агрегатного значення маркеру представленості $Aggregate^{n}_{i=1} Mrk_i$ та коефіцієнту представленості $koef^{node}_{\text{предст}}$ виконуємо фільтрацію багатовимірної моделі згідно рішення про представленість (8) та SyncroFlg = true, згортаємо за<R, A, tup> та підраховуємо кількість запитів. Співвідношення отриманого значення до загальної кількості запитів згідно з (8) і буде показником рівня необхідності синхронізації даних

$$F_{synchro} = \frac{p^{modif}_{node}}{p_{node}}, \quad (18)$$

де p^{modif}_{node} – кардинальність відношення, що включає запити віддаленого вузла (згідно рішенням про представленість), до яких входять значення атрибутів кортежів (комірки), що також входять до множини R" $^{modif}_{node}$, а p_{node} – кардинальність всіх запитів, атрибути та кортежі яких представлені у віддаленому вузлі.

Висновки

На основі реляційної моделі даних та елементів теорії множин формалізовано поняття зрізу даних множини відношень БД. Із використанням визначення операцій вибірки та зрізу, а також, враховуючи ієрархічну структуру користувацьких запитів, побудовано модель, що описує їх структуру. Дана модель, окрім відображення їх ієрархічної структури та підмножин даних і їх схем, включає аналітичні характеристики щодо типу автоматизованого робочого місця, типу(ролі) користувача, розташування та додаткових аналітичних вимірів [15]. Запропонована модель користувацького запиту, що включає ієрархічний набір результуючих відношень, кожне з яких, у свою чергу, є підмножиною базового відношення схеми БД, дозволяє проаналізувати деяку їх послідовність. Результати даного аналізу дозволяють виділити для кожного базового відношення деяку підмножину (відношення віддаленого вузла), що складатиметься із елементів, які входять до набору результуючих відношень послідовності SQL-запитів.

Введено поняття маркеру представленості даних для елементів аналітичних вимірів, на основі значення якого за допомогою запропонованої функції агрегації розраховується рівень маркеру представленості для кожного атрибуту та кортежу відношення. Для визначення оптимального значення рівня маркеру представленості введено декілька критеріїв оптимальності та побудовано математичні моделі для кожного з них, що дозволяють розрахувати їх значення в залежності від граничного рівня маркеру представленості даних на вузлі територіально розосередженої КІС.

В результаті отримано багатокритеріальну задачу, що має бути розв'язана для визначення оптимального рівня представленості даних. Слід відмітити, що критерії оптимальності, моделі яких були визначені, є незалежними та монотонними і визначаються на множині дійсних чисел в інтервалі [0;1]. Отримання оптимальних множин за Паретто та Слейтором [16] можуть дати результати лише на першому етапі моделювання, але при розрахунку оптимального рівня маркеру представленості виявляються малоефективними через зниження рівня одних критеріїв оптимальності при одночасному підвищенні інших. Простір можливих рішень визначається на множині дійсних чисел, а отже множина рішень містить велику кількість альтернатив. Перспективою розвитку та наступним етапом даного дослідження є розв'язання задачі багатокритеріальної оптимізації на основі запропонованої та наведеної вище математичної моделі.

Список літератури

[1] M. Tamer Özsu, Patrick Valduriez, *Principles of Distributed Database Systems 3rd ed. Springler*, 2011, 845 p.

[2] М. Л. Дворецький, С. Ю. Боровльова, С. В. Дворецька, «Web-застосунок складського обліку в неавтоматизованих торгових точках», *Наукові праці: науково–методичний журнал. Комп'ютерні технології*, Миколаїв: Вид-во ЧНУ ім. П. Могили, 2019.

[3] 1С:Предприятие 8. Управление торговым предприятием для Украины. [Електроний ресурс]. Режим доступу: URL: http://rarus.com.ua/torgovyy-i-skladskoy-uchet/1S_Predpriyatie_8_Upravlenie_torgovym_predpriyatiem_dlya_Ukrainy_/.

[4] Наталья Козлюк, Светлана Угримова, *Складской учет на предприятиях торговли*. Ростов-на-Дону, Россия: Феникс, 2005, 285 с.

[5] Джеффри Д. Ульман, Дженнифер Уидом, Гектор Гарсиа-Молина, *Системы баз данных. Полный курс*. Вильямс, 2004, 1088 с.

[6] Автоматическая синхронизация распределенных баз данных в разделенном режиме. [Електроний ресурс]. Режим доступу: http://stimul.kiev.ua/materialy.htm?a=avtomaticheskaya_sinkhronizatsiya_raspredelennykh_baz_dannykh_v_razdelennom_rezh.

[7] Максим Кузнецов, Игорь Симдянов, *MySQL 5*. СПб, Россия: БХВ-Петербург, 2010, 342 с.

[8] Д. Петкович, *Microsoft SQL Server 2008. Руководство для начинающих. Пер.с англ.* Спб, Россия: БХВ-Питербург, 2009, 752 с.

[9] Є. В. Малахов, Осн*ови проектування баз даних: Навч. посіб. для студ. вищих навч. закладів*. О., Україна: Наука і техніка, 2006, 156 с.

[10] В. В. Пасічник, В. А. Резніченко, *Організація баз даних та знань*. Київ, Україна: Видавнича група BHV, 2006, 384 с.

[11] Использование синхронных и асинхронных операций базы данных. [Електроний ресурс]. Режим доступу: URL: http://help.adobe.com/ru_RU/as3/dev/WS5b3cc c516d4fbf351e63e3d118666ade46-7d39.html

[12] М. Л. Дворецький, С. В. Дворецька, Є. О. Давиденко, «Інформаційна технологія визначення корисних даних при оптимізації структури та мінімізації обсягів вузла розподіленої БД», *Вісник Черкаського державного технологічного університету*, 4/2019, с. 26-35, 2019.

[13] Mykhailo Dvoretskyi, Svitlana Dvoretska, Yuriy Nezdoliy, Svitlana Borovlova, «Data Utility Assessment while Optimizing the Structure and Minimizing the Volume of a Distributed Database Node», in *Proceedings of the 1st International Workshop on Information-Communication Technologies & Embedded Systems (ICTES 2019)*, Mykolaiv, November 14-15, 2019, p. 128-137.

[14] А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод, *Методы и модели анализа данных: OLAP и Data Mining*. Питербург, Россия: БХВ-Петербург, 2004, 336 с.

[15] М. Л. Дворецький, Є.О.Давиденко, С. Ю. Боровльова, «Проектування структури розподіленої БД на базі парсингу SQL-запитів», *Наукові праці: Науково-методичний журнал. Комп'ютерні технології*, с. 26-35, 2017.

[16] І. І. Коваленко, Є. О. Давиденко, «Парето-оптимальний вибір при формуванні портфеля замовлень IT-проектів», *Наукові праці: науково-методичний журнал. Комп'ютерні технології*, т. 173, вип. 161, с. 44-48, 2011.

Стаття надійшла: 22.06.2020.

References

[1] M. Tamer Özsu, Patrick Valduriez, *Principles of Distributed Database Systems 3rd ed. Springler*, 2011, 845 p.

[2] M. L. Dvoretskyi, S. Yu. Borovlova, S. V. Dvoretska, «Web-zastosunok skladskoho obliku v neavtomatyzovanykh torhovykh tochkakh», *Naukovi pratsi: naukovo–metodychnyi zhurnal. Kompiuterni tekhnolohii*, Mykolaiv: Vyd-vo ChNU im. P. Mohyly, 2019.

[3] 1S:Predprijatie 8. Upravlenie torgovym predprijatiem dlja Ukrainy. [Elektronij resurs]. Rezhim dostupu: URL: http://rarus.com.ua/torgovyy-i-skladskoy-uchet/1S_Predpriyatie_8_Upravlenie_torgovym_predpriyatiem_dlya_Ukrainy_/.

[4] Natal'ja Kozljuk, Svetlana Ugrimova, *Skladskoj uchet na predprijatijah torgovli*. Rostov-na-Donu, Rossija: Feniks, 2005, 285 s.

[5] Dzheffri D. Ul'man, Dzhennifer Uidom, Gektor Garsia-Molina, *Sistemy baz dannyh. Polnyj kurs*. Vil'jams, 2004, 1088 s.

[6] Avtomaticheskaja sinhronizacija raspredelennyh baz dannyh v razdelennom rezhime. [Elektronij resurs]. Rezhim dostupu: http://stimul.kiev.ua/materialy.htm?a=avtomaticheskaya_sinkhronizatsiya_raspredelennykh_baz_dannykh_v_razdelennom_rezh.

[7] Maksim Kuznecov, Igor' Simdjanov, *MySQL 5*. SPb, Rossija: BHV-Peterburg, 2010, 342 s.

[8] D. Petkovich, Microsoft SQL Server 2008. *Rukovodstvo dlja nachinajushhih*. Per.s angl. Spb, Rossija: BHV-Piterburg, 2009, 752 s.
[9] Ie. V. Malakhov, Osnovy proektuvannia baz danykh: *Navch. posib. dlia stud. vyshchykh navch. zakladiv*. O., Ukraina: Nauka i tekhnika, 2006, 156 s.
[10] V. V. Pasichnyk, V. A. Reznichenko, *Orhanizatsiia baz danykh ta znan*. Kyiv, Ukraina: Vydavnycha hrupa BHV, 2006, 384 s.
[11] Ispol'zovanie sinhronnyh i asinhronnyh operacij bazy dannyh. [Elektronij resurs]. Rezhim dostupu: URL: http://help.adobe.com/ru_RU/as3/dev/WS5b3cc c516d4fbf351e63e3d118666ade46-7d39.html.
[12] M. L. Dvoretskyi, S. V. Dvoretska, Ye. O. Davydenko, «Informatsiina tekhnolohiia vyznachennia korysnykh danykh pry optymizatsii struktury ta minimizatsii obsiahiv vuzla rozpodilenoi BD», *Visnyk Cherkaskoho derzhavnoho tekhnolohichnoho universytetu*, 4/2019, s. 26-35, 2019.
[13] Mykhailo Dvoretskyi, Svitlana Dvoretska, Yuriy Nezdoliy, Svitlana Borovlova, «Data Utility Assessment while Optimizing the Structure and Minimizing the Volume of a Distributed Database Node», in *Proceedings of the 1st International Workshop on Information-Communication Technologies & Embedded Systems (ICTES 2019)*, Mykolaiv, November 14-15, 2019, p. 128-137.
[14] A. A. Barsegjan, M. S. Kuprijanov, V. V. Stepanenko, I. I. Holod, *Metody i modeli analiza dannyh: OLAP i Data Mining*. Piterburg, Rossija: BHV-Peterburg, 2004, 336 s.
[15] M. L. Dvoretskyi, Ye.O.Davydenko, S. Yu. Borovlova, «Proektuvannia struktury rozpodilenoi BD na bazi parsynhu SQL-zapytiv», Naukovi *pratsi: Naukovo-metodychnyi zhurnal. Kompiuterni tekhnolohii*, s. 26-35, 2017.
[16] I. I. Kovalenko, Ye. O. Davydenko, «Pareto-optymalnyi vybir pry formuvanni portfelia zamovlen IT-proektiv», *Naukovi pratsi: naukovo-metodychnyi zhurnal. Kompiuterni tekh*nolohii, t. 173, vyp. 161, s. 44-48, 2011.

Відомості про авторів

Фісун Микола Тихонович – доктор технічних наук, професор, завідувач кафедри інженерії програмного забезпечення Чорноморського національного університету ім. П. Могили.

Дворецький Михайло Леонідович – старший викладач кафедри інженерії програмного забезпечення Чорноморського національного університету ім. П. Могили.

Дворецька Світлана Володимирівна – старший викладач кафедри інженерії програмного забезпечення Чорноморського національного університету ім. П. Могили.

Н. Т.Фисун, М. Л. Дворецкий, С. В. Дворецкая

ПОСТРОЕНИЕ МОДЕЛИ ДЛЯ ОПТИМИЗАЦИИ СТРУКТУРЫ БАЗЫ ДАННЫХ УЗЛА В КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Черноморский национальной университет им. П. Могилы, Николаев

M. Fisun, M. Dvoretskyi, S. Dvoretskaya

BUILDING A MODEL TO OPTIMIZE THE DATABASE STRUCTURE OF THE NODE IN CORPORATE INFORMATION SYSTEMS

Petro Mohyla Black Sea National University, Mykolaiv