

ВІДГУК

офіційного опонента

доктора технічних наук, професора Філатова Валентина Олександровича
на дисертаційну роботу Дворецького Михайла Леонідовича
«Моделі та інформаційна технологія оптимізації структури бази даних вузла
у корпоративних інформаційних системах» подану на здобуття наукового
ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні
технології

Актуальність теми.

У теперішній даний час для зберігання і обробки даних активно використовуються інформаційні системи на основі реляційних баз даних, при цьому їх розмір, обсяг збережених даних, може варіюватися в широкому діапазоні - від кілобайт до терабайтів. Чим більше розмір таблиць у бази даних, тим більше часу йде на пошук та обробку потрібної інформації, збільшується час обробки запитів сервером. З одного боку цю проблему намагаються вирішувати продуктивністю комп'ютерів, на яких знаходяться системи управління базами даних, однак, набагато більшого ефекту можна досягти за допомогою зміни алгоритмів обробки запитів до баз даних.

Оптимізувати виконання запитів можна різними шляхами - за швидкістю виконання, по завантаженню процесора, за обсягом використовуваної пам'яті і т.д. Зазвичай при оптимізації запитів, головною метою є мінімізація часу виконання запиту, але оскільки при виконанні конкретного запиту найдорожчою операцією є звернення до диска, то і параметром оптимізації, як правило, є кількість звернень до диска. Іншим важливим критерієм є необхідне для обробки запиту процесорний час. Іншими складовими, як правило, нехтують.

На сьогоднішній день існує декілька способів збільшення швидкості виконання запитів - це і використання секціонованих уявлень, додаткових структур, таких як індекси і хеш-функції, використання різних алгоритмів виконання операцій з'єднання і обмеження, ведення статистики даних і багато інших.

Таким чином, хоча роботи по оптимізації запитів ведуться вже не одне десятиліття, в даний час вони не втратили актуальності. У зв'язку зі збільшенням темпів зростання обсягу інформації та навантаження на сервери баз даних, такі роботи стали ще більш актуальними.

Основний зміст роботи.

У роботі автором проведено дослідження використання розподілених та територіально розосереджених комп'ютерних систем. Шляхом реалізації інформаційної технології оптимізації структури бази даних (БД) віддаленого вузла на основі статистики Structured Query Language - запитів (SQL-запитів) розроблена та досліджена модель залежності критеріїв оптимальності структури БД віддаленого вузла розподіленої комп'ютерної

інформаційної системи (РКІС) щодо статистичних властивостей даних SQL-запитів до бази даних. Розрахунок значень здійснюється з застосуванням багатовимірної БД із виконанням відповідних зрізів даних згідно функції агрегації. Означена інформаційна технологія дає можливість виконання оперативно-аналітичного аналізу поточного стану інформаційної системи.

При побудові моделі запропоновано отримувати діапазони ключів запиту для визначення множини кортежів відношення результуючого набору рядків. Вибір найкращої альтернативи розраховується методом аналізу ієрархій із автоматичною ініціалізацією матриці попарних порівнянь альтернатив та нормалізованих значень. Розроблено інформаційну технологію розрахунку критеріїв оптимальності структури бази даних вузла розподіленої системи на базі статистики SQL-запитів. Виконано впровадження результатів роботи та отримано позитивний ефект.

У першому розділі розглянуто підходи та методи підвищення швидкості роботи інформаційних систем за рахунок оптимізації запитів. Загальним недоліком проаналізованих підходів визначено необхідність модифікації самої інформаційної системи. Розглядаючи проблеми використання «універсальних» або «комплексних» систем означеного класу, відмічається тенденція розробки окремих спеціалізованих рішень із подальшою синхронізацією даних. Розглянуто різні концепції представлення даних в розподілених БД, серед яких виділено комбіновану, що поділяє базу даних на логічні фрагменти та дозволяє мати довільну кількість фізичних копій кожного фрагмента. Ключовим фактором, який впливає на надійність і доступність бази даних, є локалізація посилань. Враховуючи закритість коду SQL-запитів, задача оптимального проектування структури бази даних зводиться до пошуку оптимального розподілу даних по вузлах розподіленої системи.

Натомість запропоновано накопичення статистичних даних виконання SQL-запитів до розподіленої БД на працюючих серверах, що дає змогу врахувати всі особливості РКІС. Парсинг тестів SQL-запитів та розщеплення їх на окремі підзапити дозволить виявити множини відношень, атрибутів та кортежів БД, задіяних у кожному запиті. Подальший аналіз накопиченої та обробленої статистики в аналітиці доступних вимірів у вигляді програмного додатку, хоста, користувача та ін. дозволить виявити частоту та характер звернень до окремих відношень, їх атрибутів і кортежів. Вводиться набір критеріїв оптимальності структури БД та формулюється задача багатокритеріальної оптимізації. Проаналізовано різні методи її розв'язання, серед яких виділяється метод аналізу ієрархій.

Зроблено висновок про доцільність досліджень, спрямованих на підвищення рівня доступності даних та ефективності використання розподілених та територіально розосереджених комп'ютерних систем шляхом реалізації інформаційної технології оптимізації структури БД вузла РКІС на базі статистики SQL-запитів.

У другому розділі на основі реляційної моделі даних та операцій реляційної алгебри реалізовано модель SQL-запиту, що включає дані про

додаток, хост, користувача, а також набір відношень, атрибутів та кортежів БД, які входять до результуючого набору кортежів. При виконанні аналізу щодо представленості даних для елементів вимірів вводиться характеристика «маркер представленості даних», що відображає рівень необхідності представлення даних на вузлі інформаційної системи. Для кожного елементу значення маркеру приймається одне з множини лінгвістичних значень {«необхідно», «бажано», «не потрібно»}, що визначає ступінь необхідності представлення даних на вузлі розподіленої системи.

Для розв'язання завдання формалізації критеріїв оптимальності граничного рівня маркеру представленості даних визначено наступні критерії оптимальності: «незалежність БД», «розмір БД» та «необхідність синхронізації», кожен з критеріїв має відповідну математичну модель, що описує його стан від значення граничного рівня маркеру представленості даних на вузлі РКІС.

При синхронізації даних запропоновано використання динамічного визначення періоду обміну даними. Наявність механізмів обчислення моменту наступної синхронізації даних дозволяє підвищити актуальність даних, та розвантажити ресурси серверів баз даних. При їх реалізації враховані наступні фактори: кількість змін у журналі транзакцій; завантаженість серверу поточними запитами.

У третьому розділі розроблено метод вибору найкращої альтернативи для граничного рівня маркеру представленості даних. Використано метод аналізу ієрархій із автоматичною ініціалізацією матриці попарних порівнянь альтернатив відповідно до отриманих математичних моделей та нормалізації значень. При побудові ієрархічної моделі процесу прийняття рішення вибору кращої альтернативи було використано 4 рівня ієрархії, а саме: «мета», «зацікавлені особи», «критерії», «альтернативи». З метою спрощення задачі було визначено п'ять альтернатив рівня маркеру представленості: «низький», «нижче середнього», «середній», «вище середнього», та «високий». Використовуючи шкалу відносної важливості критеріїв та із залученням людини, що приймає рішення, виконується процедура побудови матриці попарних порівнянь для зацікавлених осіб. Наявність математичних моделей критеріїв оптимальності дозволяє виконати розрахунок та початкову ініціалізацію матриць попарних порівнянь альтернатив на базі числових значень маркеру представленості даних для кожної альтернативи. Керуючись принципами парних порівнянь виконується нормування значень із використанням модифікованої формули нормалізації.

З метою підвищення точності результату, запропоновано представити вектор глобальних пріоритетів альтернатив у вигляді вектору нечітких чисел для маркеру представленості даних. Для отримання числового значення оптимального рівня маркеру представленості даних виконується об'єднання результатів та дефазифікація.

Оскільки на рівень представленості при аналізі користувацьких запитів впливає комбінація значень атрибутів кортежу відношення,

запропоновано представити завдання визначення рівня необхідності представлення нових даних у вигляді задачі класифікації, що розв'язується із використанням наївного Байєсівського алгоритму.

Четвертий розділ присвячено створенню інформаційної технології обліку та аналізу SQL-запитів. Для можливості отримання набору відношень, атрибутів та кортежів бази даних, було реалізовано підсистему парсингу тексту SQL-запиту. Завдання визначення множин кортежів запиту вирішено за рахунок виконання додаткового запиту для кожної таблиці із колекції відношень базового SQL-запиту. Результат запиту фіксується як множина кортежів таблиці, що використовується поточним запитом.

Враховуючи великі обсяги статистичних даних, а також необхідність виконання подальшого аналізу накопичених даних з точки зору множинності вимірів, дані SQL-запитів представлено у вигляді багатовимірної бази даних. Даний підхід підвищує прозорість структури бази даних та є своєрідним інтерфейсом при взаємодії багатовимірної моделі із реляційною. Багатовимірну базу даних реалізовано на базі SQL Analysis Server із використанням пакету SQL Server Development Tools for Business Intelligence. Структура всіх вимірів багатовимірного кубу є типовою. На базі створених вимірів та двох таблиць фактів, окремо для атрибутів та кортежів відношення, створюється два OLAP-куба, що мають в своїй основі схему «зірка» – по одній таблиці фактів та по одній таблиці на кожний вимір.

У якості клієнтського додатку в межах поточного дослідження використано MS Excel, що має механізми з підключення та отриманню даних із SQL Analysis Services та надає достатній рівень функціональних можливостей для побудови зрізів даних, необхідних для прийняття рішення щодо подальшої фільтрації даних при розрахунку значень критеріїв оптимальності структури вузла інформаційної.

У п'ятому розділі представлено програмну реалізацію інформаційної технології визначення оптимального рівня маркеру представленості даних на вузлі РКИС, що виконана у вигляді сервіс-орієнтованого WEB додатку на основі трьох-ланкової архітектури. У якості сервера БД обрано SQL Server, бекенд складова (application server) реалізована на базі фреймворку Symfony із використанням скриптової мови PHP. Взаємодія із БД реалізована із використанням Object-Relational Mapper (ORM), що дозволяє оперувати із даними за допомогою об'єктної моделі.

Зв'язок теми з планами наукових досліджень установи, де виконувалась дисертація.

Дисертація виконана у відповідності до завдань науково-дослідницьких робіт Чорноморського національного університету (надалі – ЧНУ) ім. Петра Могили, в яких автор брав участь як виконавець - «Інтеграція інформаційних технологій баз даних, баз знань та Data Mining» (номер державної реєстрації № 0112U007749); «Інтелектуалізація інформаційних систем та СППР за рахунок впровадження методів ситуаційного моделювання, сценарного аналізу, Data Mining та OLAP

технологій» (№ 0114U004402); «Розроблення найсучаснішого інтерактивного навчально-тренажерного та аналітично-консультативного комплексу військово-цивільного призначення» (№ 0118U000193).

Наукова новизна одержаних результатів.

До основних наукових результатів, що отримані автором, можна віднести наступні:

- вперше обґрунтовано та запропоновано поняття маркеру стану даних на вузлі розподіленої корпоративної інформаційної системи для елементів вимірів моделі SQL-запиту та розроблена функція агрегації, що дозволяє визначити інформаційний рівень відношення бази даних на основі статистики запитів;

- вперше розроблено інформаційну технологію розрахунку критеріїв оптимальності структури бази даних вузла розподілених корпоративних інформаційних систем, що дозволяє визначити граничний рівень маркеру даних і прийняти рішення про представленість даних на вузлі та, на відміну від існуючих, використовує аналітичні дані та результати парсингу тексту SQL-запитів до серверів розподіленої корпоративної інформаційної системи;

- удосконалено модель SQL-запиту, яка, на відміну від існуючих, представлена моделлю багатовимірної бази даних в термінах множини атрибутів та кортежів відношення, а також інших характеристик запиту, що дозволяє провести оперативно-аналітичний аналіз, зважаючи на множинність вимірів з метою розрахунку рівня представленості атрибутів та кортежів відношення бази даних на вузлі розподіленої корпоративної інформаційної системи;

- отримав подальшого розвитку метод асинхронної реплікації даних за рахунок динамічного визначення моменту наступної синхронізації даних на основі статистики SQL-запитів, що дозволяє підвищити актуальність даних та розвантажити ресурси серверів баз даних.

Обґрунтованість та достовірність наукових результатів.

Обґрунтованість та достовірність наукових результатів забезпечується коректністю постановки розглянутої задачі, строгістю використання методів системного аналізу, а також теоретичними положеннями сучасної теорії розподілених інформаційних систем та реляційних баз даних.

Значення результатів роботи для теорії та практики.

Практична значимість одержаних результатів полягає у розробленні інформаційної технології підтримки прийняття рішення при визначенні представленості даних на вузлі розподіленої корпоративної інформаційної системи. Результати дослідження впроваджено на підприємстві «Юнікс Трейд Ко» при проектуванні структури БД віддаленого вузла РКІС, що дозволило виявити закономірності між значеннями критеріїв оптимальності структури даних та рівнем представленості даних інформаційної системи на базі статистики SQL-запитів. Ефект від впровадження зазначених результатів полягає у підвищенні ефективності використання БД вузла РКІС на 12 %.

Розроблені моделі та інформаційна технологія впроваджені також на «ТОВ Еліт Білдінг», ефект від впровадження полягає у підвищенні швидкості виконання запитів на 14 % та доступності даних на 21 %. Результати роботи впроваджені в навчальний процес на кафедрі інженерії програмного забезпечення ЧНУ ім. Петра Могили при викладанні дисциплін «Інформаційні технології OLTP, OLAP та DM на клієнт-серверній платформі», «Клієнт-серверні СКБД та аналітичні системи», «Database development».

Публікація результатів дисертації.

За темою дисертації опубліковано 28 наукових робіт, з них: 17 – у фахових виданнях; 5 публікацій, що реферуються у Scopus; 6 – у збірниках матеріалів міжнародних та всеукраїнських науково-технічних конференцій.

Усі теоретичні та експериментальні дослідження виконані дисертантом самостійно.

Апробація результатів дисертації.

Основні положення і наукові результати дисертаційної роботи доповідалися і отримали позитивну оцінку на 18 міжнародних, всеукраїнських та регіональних наукових конференціях, 4 з яких входять до наукометричної бази SCOPUS.

Дисертація написана українською мовою, стиль викладу матеріалу відповідає прийнятому в науковій літературі. Зміст автореферату ідентичний змісту основних положень дисертації.

Зауваження до дисертації.

1. У першому розділі аналізуючи розподілені корпоративні інформаційні системи та методи підвищення швидкості їх функціонування, автором не наведено порівняльні графіки, таблиці, діаграми, що трохи знижує переконливість огляду завдань дослідження і формування шляхів їх вирішення.

2. У другому розділі дисертаційної роботи автор дуже докладно описує основи реляційної теорії, що не властиво для дослідження призначеного для фахівців в області інформаційних технологій, при цьому автор не дає чіткого визначення корпоративної розподіленої інформаційної системи, вузла в системі, класи такого типу об'єктів залежно від їх технічних параметрів.

3. У дисертаційній роботі не проведені дослідження ефективності запропонованого підходу для корпоративних інформаційних систем в залежності від зв'язності відносин в реляційній схемі даних, від систем з «сильно зв'язаною схемою реляційної бази даних», до схем властивим сховищам даних типу «зірочка».

4. У роботі не наведено рекомендації і порядок впровадження розробленої інформаційної технології для різних за технічними, структурним і організаційним особливостям розподіленими корпоративними інформаційними системами.

5. Дисертаційна робота не позбавлена стилістичних помилок, наприклад, «вебзастосунок» (сторінка 8) замість «WEB-додаток», «механізм

для створення» (сторінка 37) замість «алгоритм...» або «технологія створення».

Вказані вище зауваження не мають принципового значення і не зменшують наукову та практичну цінність результатів роботи.

Загальні висновки.

Дисертаційна робота Дворецького М.Л. «Моделі та інформаційна технологія оптимізації структури бази даних вузла у корпоративних інформаційних системах» є завершеним науковим дослідженням, що виконана на високому науковому рівні, в якій отримано нові науково обґрунтовані та апробовані результати.

Робота відповідає паспорту спеціальності 05.13.06 – інформаційні технології. За актуальністю розглянутих задач, обсягом досліджень, науковим рівнем і практичною цінністю отриманих результатів дисертаційна робота повністю відповідає вимогам пп. 9, 11-13 «Порядку присудження наукових ступенів і присвоєння вченого звання старшого наукового співробітника», затвердженого постановою Кабінету Міністрів України від 24 липня 2013р. № 567 (з відповідними змінами за 2015-2016рр.) до кандидатських дисертацій.

Здобувач Дворецький Михайло Леонідович заслуговує присудження наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології.

Офіційний опонент

доктор технічних наук, професор, завідувач
кафедри штучного інтелекту Харківського
національного університету радіоелектроніки

«30» 11 2020 р



В.О. Філатов

Підпис В.О. Філатов засвідчую:

Учений секретар
Харківського національного університету
радіоелектроніки

«30» 11 2020 р.




І.В. Магдаліна