УДК 004.64 DOI: https://doi.org/10.30837/IYF.IIS.2024.625

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ГЕНЕРАЦІЇ СИНТЕЗОВАНИХ ТЕСТОВИХ ДАНИХ У РЕЛЯЦІЙНИХ БАЗАХ ДАНИХ

Башкіров М. О.
Науковий керівник – д.т.н., проф. Мінухін С. В.
Харківський національний університет радіоелектроніки, каф. СТ
м. Харків, Україна
e-mail: myroslav.bashkirov@nure.ua

The work examines effective methods for generating synthetic test data for relational database systems. Three main approaches are reviewed: cyclic generation, recursive CTEs, and temporary tables. Key criteria for comparative analysis are identified: speed, resource use, scalability, complexity. By studying the strengths and weaknesses of each method, practical recommendations for selecting optimal test data generation strategies are developed. The work will benefit software testing specialists in choosing data generation methods to improve testing quality.

Синтезовані тестові дані відіграють важливу роль у процесі тестування та налаштування продуктивності сучасних систем обробки даних. Від якості та реалістичності тестових наборів залежить ефективність виявлення дефектів бази даних, а також коректна оцінка продуктивності системи в умовах реальних навантажень. Тому вибір оптимальних методів генерації даних є одним з ключових факторів успіху процесу тестування. Метою роботи є порівняльний аналіз існуючих методів генерації синтезованих тестових даних у реляційних базах даних та визначення рекомендації щодо їх ефективного застосування в залежності від цілей тестування. Дослідження включає в себе завдання, які полягають у наданні загальної характеристики основних методів генерації тестових даних у базах даних, а також аналіз переваг та недоліків кожного з цих методів. В рамках дослідження треба розробити рекомендації щодо вибору ефективних методів генерації даних для конкретних завдань використання реляційних баз даних в умовах масштабованості. Результати дослідження дозволять приймати обґрунтовані рішення при виборі стратегій генерації тестових даних та підвищити ефективність процесів тестування баз даних.

Для аналізу використано такі базові підходи до генерування тестових даних:

– використання реальних даних з виробничих систем;
– генерування даних з допомогою програмних методів [1].

Для генерації даних в реляційних базах найчастіше застосовують програмні методи на основі моделей та розподілів. Розглянемо три основні техніки реалізації таких методів:

– циклічна генерація даних з використанням звичайних циклів та запитів INSERT;

– використання рекурсивних загальних табличних виразів (CTE);

– застосування тимчасових таблиць.

Циклічна генерація проста у реалізації, але погано масштабується на великі обсяги даних. Рекурсивні CTE забезпечують високу продуктивність і гнучкість, але вимагають більш складної реалізації. Тимчасові таблиці дозволяють економити ресурси, проте мають гіршу швидкодію ніж CTE.

В якості предметної області для порівняльного аналізу методів генерації тестових даних обрано дані та модель бази даних типової торговельної компанії [2]. Вона дозволяє сформувати достатньо репрезентативну та реалістичну вибірку даних, що відображає складні ієрархічні зв'язки між сутностями таблиц баз данях та різноманітні бізнес-процеси.

Для торговельної компанії характерні наступні ключові особливості:

– наявність великої номенклатури товарів, що класифікуються за категоріями, брендами, різними характеристиками;

– процеси оформлення замовлень клієнтами з деталізацією замовлених товарів та їх кількості;

– бізнес-процеси обробки та виконання замовлень: підтвердження, резервування, комплектація, доставка, оплата.

Отже, саме тестування даних цієї предметної області дозволяє отримати об'єктивні результати для порівняння різних методів генерування тестових наборів даних.

У дослідженні проведено порівняльний аналіз трьох основних методів генерації тестових даних на прикладі реляційної СУБД Microsoft SQL Server: циклічної генерації за допомогою операторів циклу та INSERT, генерації за допомогою рекурсивних загальних табличних виразів (Common Table Expressions, CTE) та методу з використанням тимчасових таблиць. У якості критеріїв ефективності методів використано час генерації, обсяг згенерованих даних та масштабованість.

Експериментальне тестування показало, що за критерієм швидкодії (часом генерації) рекурсивні CTE демонструють найкращі результати для усіх тестових таблиць. Генерація даних здійснювалася в діапазоні від 1 000 000 до 5 000 000 записів у різних таблицях обраної моделі бази даних. За економією місця в базі даних перспективним є тимчасові таблиці, оскільки вони зберігають лише необхідні атрибути. Рекурсивні CTE продемонстрували найкращу масштабованість, ефективно працюючи як з невеликими, так і великими обсягами даних. Натомість циклічна генерація виявилася найменш продуктивною через велику кількість окремих транзакцій з базою даних. Для даних у невеликих обсягах циклічний метод також може застосовуватися з огляду на його простоту, а для складних ієрархічних даних рекурсивні CTE є найбільш ефективним рішенням.

Таким чином, для генерації невеликих обсягів даних − до 1 000 000 записів − більш ефективним є циклічний метод. При потребі в великих обсягах складних даних − від 1 000 000 записів − краще застосовувати

рекурсивні CTE, оскільки вони забезпечують високу швидкодію та масштабованість [3]. Якщо критичним фактором є економія пам'яті та ресурсів (наприклад, для навантажувального тестування), доцільно обрати тимчасові таблиці [3]. Для генерації даних зі складними ієрархічними зв'язками краще підходять рекурсивні CTE завдяки ефективній оптимізації запитів вставки даних. При виборі оптимального методу генерації тестових даних треба враховувати такі ключові фактори, як обсяг даних, що потрібно згенерувати. У роботі [4] експериментальним шляхом доведена можливість використання у якості метрики продуктивності запитів при використанні сервісу Azure SQL Database кількість повернених записів та їх відношення до загальної кількості записів у таблицях баз даних при виконанні запитів різної складності. Тому у даній роботі отримано залежності об’ємів тестових даних від кількості записів у таблицях для оцінки ефективності створення запитів. У якості подальшого дослідження пропонується застосування підходів до автоматизації процесу генерації тестових даних для реалізації масштабованості та можливості самонавчання моделей даних [5].

Вибір оптимальної стратегії вибору методу генерації даних визначається конкретними цілями та обмеженнями використання даних. Таким чином, отримані результати дозволять зробити обґрунтований вибір ефективності методів генерації тестових даних при виконанні конкретних завдань тестування баз даних.

Список використаних джерел:

1. Alsharif A., Kapfhammer G. M., McMinn. DOMINO: Fast and effective test data generation for relational database schemas. International Conference on Software Testing, Validation And Verification (ICST 2018), (09-13 Apr 2018), Västerås, Sweden. 2018. P. 12–22.

2. Мінухін С. В., Башкіров М. О. Моделювання роботи з базами даних торгівельних компаній на хмарних платформах // Інформаційні системи та технології : матеріали 12-ої Міжнародної науково-технічної конференції. Частина 2. Молодіжна секція, Харків, (28 листопада 2023 – 01 грудня 2023 р.) / наук. ред. В.В. Безкоровайний, L. Petryshyn, З.В. Дудар, Ю.В. Міщеряков. Харків : ХНУРЕ, 2023. С. 45–47.

3. Загальні вирази таблиць (CTE) проти тимчасових таблиць в BigQuery. Medium. URL: https://medium.com/@sahaabhik9/common-table-expressions-ctes-vs-temporary-tables-in-bigquery-f6057e688f01 (дата звернення: 10.01.2024).

4. Minukhin S. Performance study of the DTU model for relational databases on the azure platform. Innovative technologies and scientific solutions for industries. 2022. No 1 (19). P. 27–39.

5. Oriol X., Teniente E., Maynou M., Nadal S.. Generating valid test data through data cloning. Future Generation Computer Systems. 2023. Vol. 144. P. 179–191.