УДК 004.04, 004.65

А.М. ГЛИБОВЕЦЬ
Національний університет «Києво-Могилянська академія», Київ, Україна,
e-mail: *a.glybovets@ukma.edu.ua.*

Д.В. ЗВАЖІЙ
Національний університет «Києво-Могилянська академія», Київ, Україна,
e-mail: *d.zvazhii@ukma.edu.ua.*

ВПРОВАДЖЕННЯ ІНДЕКСУ НА БАЗІ СУФІКСНОГО ДЕРЕВА ДЛЯ ПОШУКУ ПІДРЯДКІВ У СКБД ВЕЛИКОГО РОЗМІРУ

Анотація. Розглянуто переваги та недоліки впровадження індексу на базі суфіксного дерева для оптимізації операцій пошуку підрядків у СКБД у процесі роботи з даними великого розміру. Наведено теоретичні характеристики складності операцій для суфіксних дерев. Експериментально оцінено часову складність операцій пошуку підрядків для суфіксних дерев та СКБД, таких як Elasticsearch, PostgreSQL, MySQL, ClickHouse. На основі отриманих результатів підтверджено гіпотезу про потенційну ефективність впровадження індексу на базі суфіксних дерев для оптимізації операцій пошуку підрядків у СКБД.

Ключові слова: суфіксне дерево, індекс, пошук рядків, Elasticsearch, PostgreSQL, MySQL, ClickHouse.

ВСТУП

Постійне зростання запиту на діджиталізацію процесів та сервісів зумовлює дуже швидке накопичення даних. У таких умовах важливу роль відіграють системи та рішення, пов'язані з керуванням даними, адже від них залежить швидкість та ефективність виконання операцій вставки, аналізу та пошуку інформації [1, 2].

Одним з класичних підходів до покращення операцій пошуку в СКБД є використання різноманітних індексів. Це дає змогу зменшити кількість операцій порівняння, проте накладає додаткові вимоги до підтримки актуальності такого індексу, особливо для систем зі значною часткою операцій запису [3]. Важливим серед таких підходів до індексації є Б-дерево (*B*-tree) [4]. Ця структура є типовим індексом для PostgreSQL. Її також інколи застосовують для операцій пошуку повного збігу рядків та для діапазонних запитів. Щодо операцій повнотекстового пошуку та зіставлення шаблонів (pattern matching), то для них використання Б-дерева уже не є таким ефективним. У цих випадках індекс на базі *n*-gram [5] буде значно ефективнішим для операцій повнотекстового пошуку. Загалом вибір того чи іншого типу індексування сильно залежить від типу задачі та операцій, які потребують оптимізації.

Серед менш поширених підходів до індексування варто виокремити суфіксні дерева [6]. Це давно відома та широко досліджувана структура даних, яка є достатньо ефективною для операцій пошуку підрядків. Водночас використання суфіксних дерев як індексів не є поширеним підходом для сучасних СКБД.

У цій роботі досліджується можливість впровадження індексу на основі суфіксного дерева в сучасні реляційні та нереляційні бази даних для оптимізації операцій пошуку підрядків.