

WYDZIAŁ GEOLOGII, GEOFIZYKI I OCHRONY ŚRODOWISKA



AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

*Analiza porównawcza wydajności złączeń i
zagnieźdżeń dla schematów
znormalizowanych i zdenormalizowanych*

Autor:

Kierunek studiów:

Piotr Życki

Geoinformatyka

Kraków, 2023

Wstęp

Niniejsze opracowanie ma na celu porównanie wydajności wykonywania złączeń i zagnieżdżeń w różnych oprogramowaniach bazodanowych, z wykorzystaniem schematów znormalizowanych i zdenormalizowanych. Porównaniu podlega również indeksacja i jej wpływ na szybkość wykonywania poleceń. W testach posłużono się systemem PostgreSQL oraz MySQL.

Testowane tabele

W celu wykonania testów, utworzono bazę danych odzwierciedlającą schemat tabeli geochronologicznej obrazującej przebieg historii Ziemi. Baza ta powstała w dwóch wariantach: znormalizowanej do trzeciej postaci normalnej (schemat płątka śniegu) oraz zdenormalizowanej (schemat gwiazdy). Postać zdenormalizowana powstała po przez połączenie wszystkich tabel wariantu znormalizowanego w jedną – dokonano tego z wykorzystaniem polecenia:

```
CREATE TABLE geochronologia.GeoTabela AS
    (SELECT * FROM geochronologia.GeoPietro
        NATURAL JOIN geochronologia.GeoEpoka
        NATURAL JOIN geochronologia.GeoOkres
        NATURAL JOIN geochronologia.GeoEra
        NATURAL JOIN geochronologia.GeoEon );
```

Utworzono tabelę *Milion*, wypełnioną milionem rekordów, w której kolumna *liczba* stanowi kolejne liczby naturalne z przedziału od 0 do 999 999. Do jej stworzenia posłużono się tabelą *Dziesiec*, której rekordy stanowią cyfry z zakresu 0-9. Wypełnienie tabeli *Milion* rekordami było możliwe dzięki poleceniu:

```
INSERT INTO geochronologia.Milion
SELECT a1.cyfra + 10 * a2.cyfra + 100 * a3.cyfra + 1000 *
a4.cyfra + 10000 * a5.cyfra + 10000 * a6.cyfra AS liczba,
a1.cyfra AS cyfra, a1.bit AS bit
FROM
geochronologia.Dziesiec a1, geochronologia.Dziesiec a2,
geochronologia.Dziesiec a3, geochronologia.Dziesiec a4,
geochronologia.Dziesiec a5, geochronologia.Dziesiec a6;
```

Konfiguracja sprzętowa i programowa

Wszystkie testy wykonane na potrzeby tego sprawozdania odbyły się z wykorzystaniem komputera o następującej specyfikacji sprzętowej:

- CPU: Intel Core i5-10300H 2.5 GHz,
- GPU: NVIDIA GeForce GTX 1650,
- RAM: 16 GB,
- Dysk twardy: SSD 500GB,
- OS: Windows 10 Home.

Testów dokonano korzystając z oprogramowania:

- MySQL Community Server 8.0.33,
- PostgreSQL 15.3-1.

Testy wydajności

Testy złączeń i zagnieżdżeń zostały dokonane z wykorzystaniem polecenia SELECT wywołanego w 4 oddzielnych zapytaniach. Każde z zapytań uruchomiono 10-krotnie, z pominięciem pomiarów skrajnych, a wyniki wydajności stanowią wartości: średnia i minimalna, z czasów wykonywania każdego z testów. Wszystkie testy zostały przeprowadzone w dwóch etapach: pierwszym bez nałożonych indeksów na kolumny, oraz w drugim, z nałożoną indeksacją wszystkich kolumn tabel.

Zapytanie 1

W zapytaniu dokonywane jest złączenie tablicy Milion z tabelą geochronologiczną w postaci zdenormalizowanej:

```
SELECT COUNT(*) FROM geochronologia.Milion
INNER JOIN geochronologia.GeoTabela
ON
(mod(geochronologia.Milion.liczba,68)=(geochronologia.GeoTabela.id_pietro));
```

Zapytanie 2

Złączeniu podlega tablica Milion z tabelą geochronologiczną w postaci znormalizowanej, złożoną z pięciu połączonych ze sobą tabel:

```
SELECT COUNT(*) FROM geochronologia.Milion
INNER JOIN geochronologia.GeoPietro
ON
(mod(geochronologia.Milion.liczba,68)=geochronologia.GeoPietro.id_pietro)
NATURAL JOIN geochronologia.GeoEpoka
NATURAL JOIN geochronologia.GeoOkres
NATURAL JOIN geochronologia.GeoEra
NATURAL JOIN geochronologia.GeoEon;
```

Zapytanie 3

Celem tego zapytania jest złączenie tablicy Milion ze zdenormalizowaną tabelą geochronologiczną, gdzie złączenie odbywa się z wykorzystaniem zagnieżdżenia skorelowanego:

```
SELECT COUNT(*) FROM geochronologia.Milion
WHERE mod(geochronologia.Milion.liczba,68)=
(SELECT id_pietro FROM geochronologia.GeoTabela
WHERE mod(geochronologia.Milion.liczba,68)=(id_pietro));
```

Zapytanie 4

Tablica miliona wyników podlega złączeniu z tabelą geochronologiczną w postaci znormalizowanej, natomiast odbywa się to ponownie poprzez zagnieżdzenie skorelowane; w zapytaniu wewnętrznym następuje złączenie poszczególnych tablic tabeli geochronologicznej:

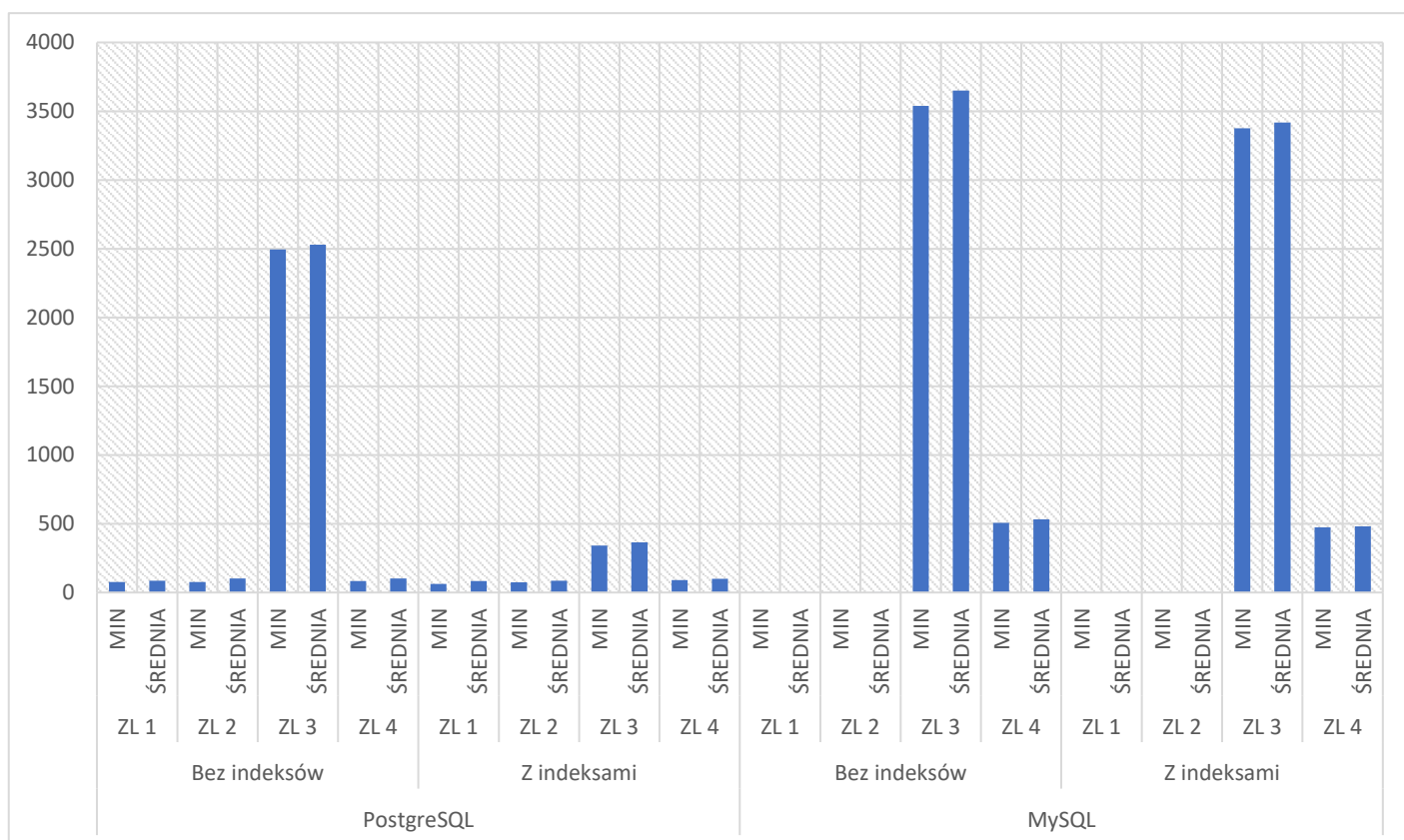
```
SELECT COUNT(*) FROM geochronologia.Milion
WHERE mod(geochronologia.Milion.liczba,68)=
(SELECT geochronologia.GeoPietro.id_pietro FROM
geochronologia.GeoPietro
NATURAL JOIN geochronologia.GeoEpoka
NATURAL JOIN geochronologia.GeoOkres
NATURAL JOIN geochronologia.GeoEra
NATURAL JOIN geochronologia.GeoEon);
```

Wyniki testów

Czasy wykonywania zapytań, wyrażone w milisekundach [ms], zostały zamieszczone w tabeli (Tabela 1. Średnie i minimalne czasy wykonywania zapytań w systemach PostgreSQL i MySQL) oraz przedstawione na wykresie (Rysunek 1).

Tabela 1. Średnie i minimalne czasy wykonywania zapytań w systemach PostgreSQL i MySQL

		Zapytanie 1		Zapytanie 2		Zapytanie 3		Zapytanie 4	
		MIN	ŚREDNIA	MIN	ŚREDNIA	MIN	ŚREDNIA	MIN	ŚREDNIA
Bez indeksów	PostgreSQL	76	85	76	102	2494	2529	84	103
	MySQL	0.33	0.37	0.46	0.54	3539.46	3650.15	506.05	532.41
Z indeksami	PostgreSQL	63	84	75	87	342	364	90	100
	MySQL	0.3	0.35	0.46	0.5	3375.03	3418.4	475.22	480.51



Rysunek 1. Porównanie czasów wykonywania poszczególnych zapytań

Wnioski

Zapytania ZL1 i ZL2 charakteryzują się w obu systemach niewielkim czasem wykonywania – a w przypadku MySQL, czasem poniżej 1 ms. Pozostałe dwa zapytania potrzebowały zdecydowanie więcej milisekund na wykonanie złączeń; wartości w tabeli wskazują, że system PostgreSQL poradził sobie z nimi ewidentnie lepiej.

Dla wszystkich zapytań, indeksacja pozwoliła na skrócenie czasu wykonywania polecenia – zmiana ta jest szczególnie widoczna w systemie PostgreSQL, gdzie dla zapytania ZL3 różnica średnich czasów wyniosła ponad 2000 ms. Dodanie indeksów w MySQL spowodowało zauważalnie mniejszą zmianę w średnich czasach dla poszczególnych poleceń, niż było to widoczne w przypadku pierwszego systemu – największa zmiana ma miejsce przy zapytaniu ZL3, osiągając wartość 232 ms.

Przyrównując czasy wykonywania zapytań ZL1 i ZL2, wykorzystujących odpowiednio: tabelę zdenormalizowaną i znormalizowaną, pośród testów dla obu systemów występuje zauważalna różnica. Zapytanie 1 jest wykonywane za każdym razem szybciej od zapytania 2, niezależnie od systemu i indeksacji, co pozwala stwierdzić, że operacja złączania tabel jest szybsza dla tych w postaci zdenormalizowanej. W przypadku zapytań wykorzystujących zagnieżdżenie skorelowane, dłuższe czasy były uzyskiwane przez zapytanie ZL3, również korzystającego z tabeli zdenormalizowanej. Operacja ta wykonywana jest krócej, gdy zastosowana zostanie tabela znormalizowana (ZL4).