Analiza wydajności baz relacyjnych i nierelacyjnych.

Wstęp:

Relacyjne bazy danych są najbardziej pożądanym przez programistów modelem od jego rozwinięcia przez Edgara Coodda w 1970 roku. Wielu profesjonalistów z branży IT zdobyło doświadczenie w zarządzaniu nimi oraz z jej wykorzystaniem powstało dużo projektów.

Kluczem jest wysoka wydajność baz relacyjnych.

Stworzenie optymalnych rozwiązań pozwalających efektywnie zarządzać danymi wymaga od projektantów i programistów znajomości możliwości baz danych. W celu ułatwienia dokonania wyboru z pośród wielu dostępnych rozwiązań warto skorzystać z gotowych porównań wydajności baz danych. Jednak, gdy aplikacja, która jest tworzona wymaga użycia specyficznych narzędzi lub typu danych niezbędne jest skorzystanie z autorskiego programu do analizy.

Czas wykonania operacji z wykorzystaniem baz danych ma znaczący wpływ na wydajność całego programu, dlatego tez niezbędne jest wybranie właściwego narzędzia do tego celu.

W tej pracy zostanie porównane relacyjne i nierelacyjne bazy danych z wykorzystaniem autorskiego programu pozwalającego na zbadanie czasu wykonywania operacji zapisu, … i usuwanie danych z wymienionych baz.

Przegląd literaturowy:

Przed przystąpieniem do badań przeprowadzono przeanalizowaną literaturę o tematyce badania wydajności różnych baz danych.

W artykule autorstwa Yusufa Abubakara została zbadana wydajność następujących baz danych: MySQL, PostgreSQL, MariaDB i SQLite [1]. W tym celu zmierzono czas wykonania operacji odczytu, zapisu oraz aktualizacji danych. Rezultatem wykonanych pomiarów było ustalenie, że największa wydajność spośród porównywanych baz danych uzyskał PostgreSQL. Rozróżniając wyniki ze względu na typ przeprowadzanej operacji wyciągnięto wniosek, że dla operacji zapisu oraz aktualizacji danych największą wydajnością wykazał się PostgreSQL, natomiast dla operacji odczytu – MariaDB.

W pracy Katarzyny Krocz, Oleksandry Kizun i Marii Skublewskiej-Paszkowskiej została dokonana analiza porównawcza wydajności relacyjnych baz danych MySQL, PostgreSQL, MariaDB oraz H2 [5]. Zbadano w niej prędkość wykonywania prostych oraz bardziej złożonych zapytań SQL, takich jak przeprowadzenie operacji dodania, aktualizacji, usuwania i wybierania danych. Po przeprowadzeniu badań wyciągnięto wniosek, że MySQL charakteryzuje się niską wydajnością dla przetwarzania dużych ilości danych. Kolejnym rezultatem analizy było zaobserwowanie, że dla bazy PostgreSQL wykonanie operacji wybierania, złączenia oraz usuwania danych zajmuje mniej czasu niż przy wykorzystaniu baz MySQL lub MariaDB, a także, że w przypadku operacji aktualizacji najkrótszy czas jej wykonania uzyskał PostgreSQL.

Artukuł Sasalaka Tongkawa i Aumnaat Tongkawa opisuje eksperyment, w którym dokonano porównania baz MariaDB oraz MySQL [2]. Jego wynikiem było dojście do wniosku, że MySQL wykazuje większą wydajność niż MariaDB.

[1] Y. Abubakar, Benchmarking popular open source RDBMS: a performance evaluation for it professionals, International Journal of Advanced Computer Technology (IJACT), 2014

[2] S. Tongkaw, A. Tongkaw, A Comparison of Database Performance of MariaDB and MySQL with OLTP Workload, IEEE Conference on Open Systems (ICOS), 2016.

[5] Katarzyna Krocz, Oleksandra Kizun, Maria Skublewska-Paszkowska, Analiza porównawcza wydajności relacyjnych baz danych MySQL, PostgreSQL, MariaDB oraz H2, 2019

[3] Shuxin, Y. and Indrakshi, R.. (2005) Relational database operations modeling with UML, Proceedings of the 19th International Conference on Advanced Information etworking and Applications, 927-932.

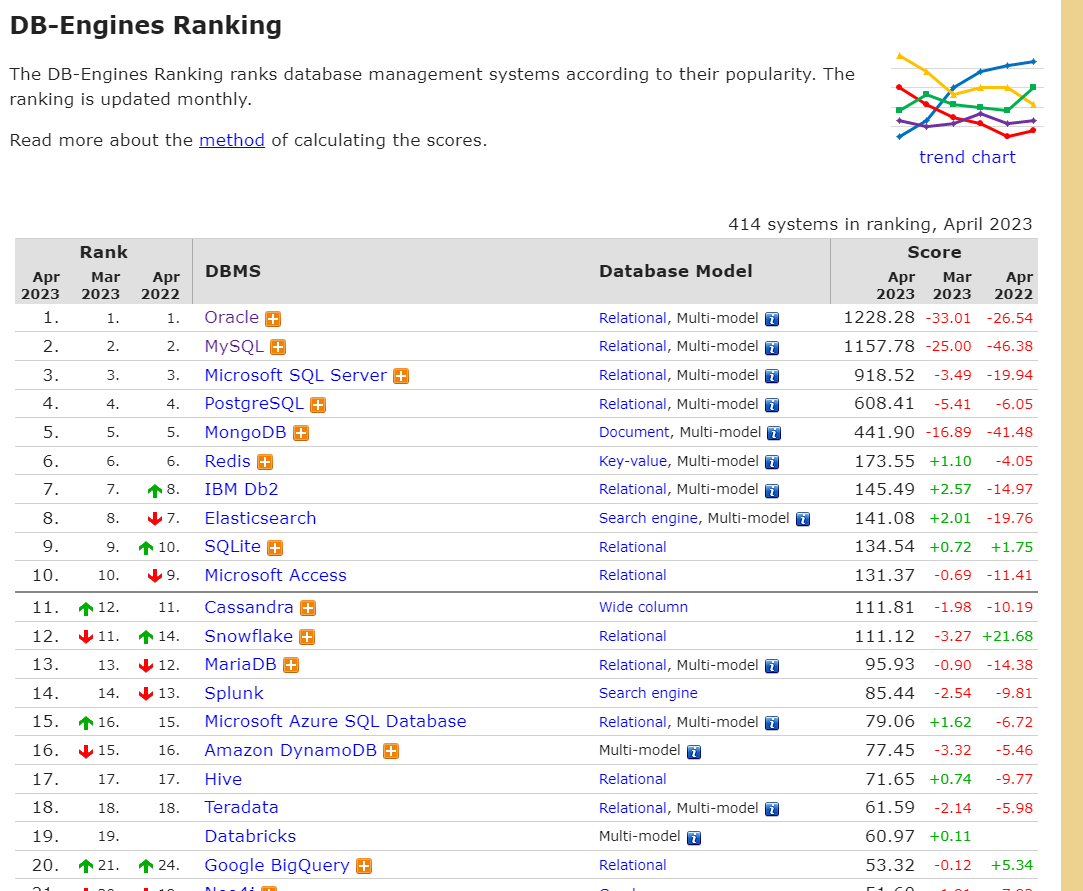
[4] Edgar F. Codd, “A relational model of data for large shared data banks”, Communications of ACM, Vol.13, No.6, June 1970, pp.377-387.

Badane bazy danych:

MySQL jest najpopularniejszym [5] opensource’owym systemem zarządzania relacyjnymi bazami danych rozwijanym przez firmę Oracle.

MariaDB Jej autorem jest współtworca systemu MySQL - Monty Widenius. MariaDB jest oparta na tym samym kodzie bazowym co MySQL i dąży do utrzymania kompatybilności z jej poprzednimi wersjami [[6]](https://pl.wikipedia.org/wiki/MySQL#cite_note-5).

[5] Ranking SZBD według ich popularności, https://db-engines.co m/en/ranking, [01.04.2023].



[6] https://mariadb.com/kb/en/mariadb-vs-mysql-features/

Dane testowe:

Do badania wydajności baz danych użyto danych z Internet Movie Database (IMDB) z dnia 28.02.2023 roku. Internet Movie Database to  największa na świecie [internetowa](https://pl.wikipedia.org/wiki/Internet) [baza danych](https://pl.wikipedia.org/wiki/Baza_danych) na temat [filmów](https://pl.wikipedia.org/wiki/Film) i ludzi z nimi związanych. Powstała 17 października 1990 roku a jej autorem jest Colin Needham. Aktualnym właścicielem jest firma [Amazon.com](https://pl.wikipedia.org/wiki/Amazon.com). [7]

Oryginalnie zbiór danych składa się z 5 plików:

- który zawiera….

[7] <https://pl.wikipedia.org/wiki/IMDb> [1.04.2023]

Do tego badania zdecydowano się wykorzystać tych 3 z plików ….. Następnie konieczne było podzielenie plików na mniejsze zbory, tak aby można było testować zmianę wydajności spowodowana wykonywaniem operacji na coraz to większej ilości danych. Do tego celu uzyto programu RapidMiner.

Przy podziale na mniejsze zbiory istotne było zachowanie spójności danych, tak by zachować połączenia między poszczególnymi wierszami. Dlatego tez podziału dokonano bazując na pliku akas…, z którego kolejno wylosowano 1000, 1000 .,… elementów, a następnie przeszukano każdy z plikow basic oraz reatings, w celu znalezienia w nich rekordów powiązanych z tymi wylosowanymi z pliku akas. Dzięki temu zapewniono istnienie połączeń między tabelami.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Akas | 1000 | 10000 | 100000 | 1000000 |
| Basics | 1000 | 9993 |  |  |
| ratings | 115 | 1179 |  | 120603 |
|  |  |  |  |  |