Wykorzystanie bazy grafowej Neo4j

Termin oddania: mocno sugerowany - poniedziałek 17.01; ostateczny – poniedziałek 24.01

Kara za przekroczenie terminu: 0 punktów za dostarczenie pracy po 24.01, w innym przypadku brak kary punktowej. Prace będą sprawdzane zaczynając od terminowych, a potem w kolejności złożenia – więc im później złożona praca, tym późniejsza informacja o zaliczeniu przedmiotu i tym mniej czasu na ewentualną poprawę.

Punkty: maksymalnie 13 (uzasadnienie wyboru i złożoność obu zbiorów danych – 2, zapytania – 2, indeksy i optymalizacja zapytań – 2, funkcje przestrzenne - 2, procedura – 1, analiza zbioru danych, modyfikacje i wnioski – 2, APOC - 2);

Wymagania minimalne: Zainstalowana i skonfigurowana baza Neo4j zawierająca wczytane dane (**co najmniej 100 połączonych węzłów 5 rodzajów, co najmniej 5 rodzajów krawędzi**), skrypty umożliwiające odtworzenie całej pracy na innej maszynie z zainstalowanym Neo4j, **co najmniej 2 nietrywialne zapytania**, utworzony indeks i **oceniona jego skuteczność**, **stworzona i omówiona procedur**a.

Zakres:

1) Wybierz zbiór danych, który wykorzystasz podczas projektu i uzasadnij wybór. Zbiór musi być odpowiednich rozmiarów (**co najmniej 100 połączonych węzłów 5 rodzajów**, **co najmniej 5 rodzajów krawędzi**, jest to wymaganie minimalne, w celu otrzymania punktów zbiory powinny być znacząco większe i bardziej skomplikowane).
   Wybrany zbiór danych wpisz na MS Teams wraz z krótkim opisem, linkiem i informacjami dotyczącymi rozszerzania (**więcej o tym w punkcie 3**). Mile widziane jest wykorzystanie zbioru danych dotyczącego problematyki, która nie została jeszcze przez nikogo wybrana. Premiowane jest wykorzystanie zbioru innego niż występował w poprzednim ćwiczeniu, ale można skorzystać ze zbioru danych wykorzystanego do MongoDB i takiego, który został wykorzystany już przez innego studenta.
   Te same informacje umieść również w sprawozdaniu.
2) Zainstaluj i skonfiguruj Neo4j, zaimportuj i przetwórz potrzebne dane. W razie potrzeby utwórz potrzebne relacje. **Nie dokumentuj całej instalacji** – przedstaw tylko problemy i rozwiązania (jeśli będą) oraz rezultat (prezentując na przykład interesujący wycinek zbioru danych i go omawiając).
3) Dołącz drugi, zbliżony zbiór danych i powiąż go z pierwszym zbiorem danych (przykładowo – jeśli nasz pierwszy zbiór danych dotyczył postaci i miejsc w Grze o Tron to drugi może być na przykład mapą Westeros, zbiorem danych o aktorach i filmach czy informacjami o stoczonych bitwach). Można wykorzystać istniejący zbiór danych lub stworzyć go samemu. Umieść informację o tym, w jaki sposób planujesz rozszerzyć pierwszy zbiór danych i do czego wykorzystać – **zaprezentuj to przy pomocy zestawu co najmniej 5 zapytań** (**złożoność zapytań jest istotna dla oceny**, co najmniej jedno powinno korzystać z UNION, co najmniej jedno powinno korzystać z MERGE). Jeśli oba zbiory były istniejące utwórz i powiąż z oboma zbiorami co najmniej 5 nowych, stworzonych przez siebie, węzłów (**udokumentuj i wyjaśnij powiązania**).
4) Dla przygotowanych zapytań stwórz indeksy – przedstaw zasadę działania indeksów w Neo4j, **oceń w praktyce ich skuteczność**. Omów jakie jeszcze indeksy warto stworzyć w wykorzystywanym zbiorze danych i dlaczego. Zapoznaj się z mechanizmami optymalizacji

zapytań w Neo4j, **przedstaw praktycznie**, jak przygotowane przez ciebie zapytania można wykonać szybciej.

5) Przygotuj (jeśli nie ma) zestaw danych przestrzennych dla swojego projektu – **przedstaw w praktyce wykorzystanie, wady i zalety** funkcji przestrzennych w Neo4j. **Wykonaj co najmniej jedno zapytanie** wykorzystujące agregacje oraz co najmniej jedno korzystające z funkcji odległości (np. najkrótsza ścieżka), omów ich działanie i wydajność. Premiowane jest skorzystanie z większej liczby funkcjonalności przestrzennych, jeśli to możliwe dla wybranych zbiorów danych.
6) Stwórz co najmniej jedną procedurę i wywołaj ją przy pomocy CALL. Omów w jaki sposób procedury w Neo4j różnią się w stosunku do relacyjnych baz danych oraz co czego możemy je zastosować.
7) **Przeanalizuj końcowy zbiór danych** – oceń w jaki sposób warto byłoby go umieścić na kilku fizycznych maszynach w celu uzyskania maksymalnej wydajności, znajdź mosty i węzły przegubowe, wyszczególnij podgrafy. Omów jaki mają wpływ na zbiór danych w projekcie. Dla pełnej punktacji zaproponuj i wdróż potrzebne modyfikacje, pokaż różnice jakie spowodowały.
8) Zapoznaj się z biblioteką APOC (https://neo4j.com/labs/apoc/) – wybierz i omów najciekawsze możliwości jakie daje, wybór uzasadnij.

Pamiętaj o przygotowaniu i załączeniu wraz ze sprawozdaniem skryptów, które umożliwią załadowanie obu zbiorów danych i odtworzenie całej wykonanej pracy.

Przed przystąpieniem do ćwiczenia warto przejrzeć przykładowe zbiory danych i zapytania z nimi powiązane na https://neo4j.com/developer/example-data/

Poprawa

Indywidualne ustalenie zakresu poprawy – od 17.01 do 26.01, kontakt na czacie przez MS Teams z prowadzącym. Polecam kontaktować się jak najszybciej, czas na poprawę będzie do 28.01.