Praktikum 3: Performance-Analyse

Ziel des Praktikums

Ziel des dritten Praktikums ist es, mit größeren Datenmengen zu arbeiten und Optimierungsmöglichkeiten der Dokumentenorientierten Datenbanksysteme zu testen.

Außerdem soll die Variante der *referentiellen* Datenmodellierung der *embedded* Variante gegenübergestellt und verglichen werden.

Vorbereitung vor(!) dem Praktikumstermin

Wie in jedem DBMS, können Sie sich auch in MongoDB und Couchbase die Ausführungspläne der Anfragen mit Hilfe der jeweiligen `explain`-Kommandos ansehen. Erarbeiten Sie sich in den jeweiligen Dokumentationen die Möglichkeiten dieser Kommandos und machen Sie sich außerdem mit den Möglichkeiten zur Zeitmessung / Profiling von Anfragen in den verschiedenen DBMS vertraut. Für MongoDB können Sie zusätzlich auch das Tool `MongoDB Compass` verwenden (siehe Wiki).

Löschen Sie Ihre in Praktikum 1 bzw. Praktikum 2 eingefügten Daten und fügen Sie jeweils die 20M-Datensätze in MongoDB bzw. Couchbase ein. Machen Sie dies unbedingt **vor(!) dem Praktikumstermin**, da das Laden der Daten ggf. einige Zeit dauert.

Aufgabe 1 (Datenbankanfragen und Optimierung)

Führen Sie die untenstehenden Datenbankanfragen in beiden Datenbanksystemen aus und messen Sie die benötigte Zeit. Führen Sie die Abfragen dazu mindestens 4 Mal aus. Dokumentieren Sie, welche Vorgehensweise Sie für das Messen der Zeit im jeweiligen DBMS gewählt haben.

1. Ausgabe der Titel aller Filme, deren Titel "Matrix" enthält.
2. Ausgabe der Ratings des Films mit der id / movieId = 6365.
3. Ausgabe des Titels und der Ratings des Films mit der id / movieId = 6365.

Führen Sie die Datenbankanfragen zunächst ohne spezielle Optimierungsmaßnahmen durch.

Überlegen Sie sich dann, welche Möglichkeiten zur **Optimierung** der Anfragen (beispielsweise benutzerdefinierte Index-Strukturen o.ä.) es in den jeweiligen Datenbanksystemen gibt. Testen Sie diese und dokumentieren Sie die Ergebnisse geeignet.

Hinweise zur Dokumentation

Dokumentieren und vergleichen Sie die gemessenen Zeiten geeignet im Praktikumsbericht. Erzeugen Sie Diagramme zur Auswertung. Achten Sie dabei auf eine sinnvolle Wahl der angegebenen Größeneinheiten.

Stellen Sie im Praktikumsbericht auch dar, wie Sie bei den Messungen vorgegangen sind (Ermittlung der Zeiten, Anzahl Wiederholungen, Mittelwert, Standardabweichung etc.)

Aufgabe 2 (embedded vs. referentiell)

Vergleichen Sie am Beispiel von **MongoDB** die *referentielle* Datenmodellierung mit der *embedded* Variante. Nutzen Sie für die referentielle Variante die aufbereiteten Datensätze im Verzeichnis `../JSONref`.

Testen Sie die beiden Szenarien im Vergleich wiederum mit den obigen Anfragen.

Welche Unterschiede, Vor- und Nachteile weisen die beiden Modellierungsvarianten auf (Performance, Komplexität der Anfragen etc.)? Begründen Sie die ggf. festgestellten Performance-Unterschiede. Testen Sie auch hier ggf. geeignete Optimierungsmaßnahmen. Dokumentieren Sie Ihr Vorgehen und die Ergebnisse.

Aufgabe 3 (Zusatzaufgabe)

Führen Sie den Vergleich *embedded* vs. *referentiell* nun für **Couchbase** durch. Da der (referentielle) 20M-Datensatz sehr groß wird in Couchbase (> 7 GB) und folglich das Anlegen von Indexen etc. sehr lange dauert, arbeiten Sie in dieser Aufgabe mit dem 1M-Datensatz (aufbereitete Daten in `../couchbaseRef` – die gesuchte `movieId` lautet in diesem Datensatz `2571`).

Es geht also in dieser Aufgabe nicht vorranging um den Aspekt Performance, sondern um das Handling der Daten (Anfragen, Umgang mit verschieden strukturierten Daten im gleichen Bucket etc.) Überlegen Sie sich, ob Sie die Daten nach dem Laden in die Datenbank ggf. noch geeignet verändern sollten.

Praktikumsbericht

Abgabe des Praktikumsberichts wiederum eine Woche nach dem jeweiligen Praktikumstermin.