

Ausschreibung Abschlussarbeit

Thema:	Entwicklung eines Benchmarks zur Evaluation von Schemaoptimierungen bei der Migration zwischen relationalen und NoSQL-Datenbanksystemen
Kategorie:	Masterarbeit
Lehrgebiet:	Datenbanken und Informationssysteme, FernUniversität in Hagen
Betreuung:	André Conrad
Kontakt:	andre.conrad@fernuni-hagen.de

Themenbeschreibung

Die Migration zwischen verschiedenen Datenbanksystemen (DBS) ist ein relevantes Forschungs- und Praxisproblem. Dies betrifft insbesondere Migrationen zwischen heterogenen Datenmodellen (z.B. von PostgreSQL zu MongoDB). Während relationale Datenbanksysteme üblicherweise auf normalisierten Schemata und deklarativen Anfragesprachen wie SQL basieren, weisen NoSQL-Datenbanksysteme unterschiedliche, auf bestimmte Anwendungsfälle spezialisierte, Datenmodelle und Anfrageparadigmen auf. Dadurch führt eine direkte Übertragung eines bestehenden Schemas in vielen Fällen zu unerwarteten Performanceeinbußen und ist daher nicht sinnvoll.

Ein zentraler Bestandteil solcher Migrationen ist die Schemaoptimierung. Ziel ist es, ein Schema für das Zieldatenbanksystem zu erzeugen, das sowohl die Charakteristiken der Quelldaten als auch die Anfragen (lesende und schreibende sowie deren relative Häufigkeit) des Anwendungsszenarios berücksichtigt.

In der Forschung existieren bereits verschiedene Ansätze und Algorithmen zur Schemaoptimierung von NoSQL Datenbanksystemen. Diese Arbeiten evaluieren ihre Verfahren häufig anhand selbst definierter Benchmarkszenarien, Datensätze und Workloads. Ein Problem besteht jedoch darin, dass diese Evaluierungen nur eingeschränkt vergleichbar und reproduzierbar sind. Häufig werden Benchmarkszenarien zwischen Veröffentlichungen verändert, Datensätze sind nicht dauerhaft verfügbar oder zentrale Eigenschaften der Daten und Workloads werden nicht hinreichend präzise beschrieben. Für eine systematische Evaluation von Schemaoptimierungsverfahren wäre daher ein klar definierter und reproduzierbarer Benchmark hilfreich, ähnlich wie etablierte Benchmarks der Transaction Processing Performance Council, z. B. TPC-C, TPC-H oder TPC-DS, strukturierte Vorgaben für Daten, Workloads und Messmethodik machen.

Ziel dieser Masterarbeit ist die Konzeption eines "standardisierten" Benchmarks zur Evaluation von Schemaoptimierungen im Kontext einer Schema- und Datenmigration zwischen relationalen und NoSQL-Datenbanksystemen. Der Benchmark soll es ermöglichen, unterschiedliche Schemaoptimierungsansätze nachvollziehbar, vergleichbar und reproduzierbar zu bewerten.

Im Rahmen der Arbeit sollen zunächst existierende Benchmarks aus dem Datenbankbereich recherchiert und systematisch analysiert werden. Dazu zählen insbesondere relationale Benchmarks, NoSQL-Benchmarks sowie Benchmarks für hybride oder polyglotte Datenhaltung. Auf dieser Grundlage soll untersucht werden, welche Elemente solcher Benchmarks für die Evaluation von Schemaoptimierungen geeignet sind und welche Anforderungen ein migrationsspezifischer Benchmark für unterschiedliche NoSQL Datenbanksysteme zusätzlich erfüllen muss.

Darauf aufbauend soll ein Benchmark-Konzept entwickelt werden. Dieses soll insbesondere folgende Aspekte berücksichtigen:

- ein oder mehrere fachlich motivierte Anwendungsszenarien,
- ein relationales Ausgangsschema,
- definierte Zielmodelle für ausgewählte NoSQL-Datenbanksysteme,
- reproduzierbare Datengenerierung mit dokumentierten Datencharakteristiken,
- definierte Workloads aus Lese- und Schreiboperationen,
- Beschreibung typischer Zugriffsmuster und Anfrageklassen,
- Metriken zur Bewertung von Schemaoptimierungen,
- Vorgaben zur experimentellen Durchführung,
- Anforderungen an Dokumentation, Erweiterbarkeit und Reproduzierbarkeit.

Zielsetzung

Ziel der Arbeit ist die Entwicklung eines Benchmark-Konzepts, mit dessen Hilfe verschiedene Schemaoptimierungsverfahren systematisch evaluiert werden können.

Die Arbeit soll insbesondere folgende Forschungsfragen adressieren:

1. Welche bestehenden Benchmarks aus dem relationalen, NoSQL- und polyglotten Datenbankbereich sind für die Evaluation von Schemaoptimierungsverfahren relevant?
2. Welche Eigenschaften müssen Daten, Schemata und Workloads besitzen, um Schemaoptimierungen realistisch und vergleichbar bewerten zu können?
3. Welche Metriken eignen sich zur Bewertung optimierter Zielschemata in unterschiedlichen Datenbanksystemen?
4. Wie kann ein Benchmark so spezifiziert werden, dass Experimente leicht reproduzierbar und Ergebnisse zwischen unterschiedlichen Arbeiten vergleichbar sind?

Erwartete Ergebnisse

Die Arbeit soll folgende Ergebnisse liefern:

- eine strukturierte Übersicht und Bewertung existierender Datenbankbenchmarks,
- eine Analyse der Benchmarking-Praxis in bestehenden Arbeiten zur Schemaoptimierung,
- eine Anforderungsanalyse für einen Benchmark zur Evaluation von Schemaoptimierungen,
- eine detaillierte Benchmark-Spezifikation,
- definierte Datencharakteristiken und Workload-Klassen,
- geeignete Metriken und Auswertungsverfahren,
- eine prototypische Implementierung ausgewählter Benchmark-Bestandteile,
- eine kritische Diskussion von Grenzen und Erweiterungsmöglichkeiten des Benchmarks.

Voraussetzungen

Hilfreich sind Kenntnisse in mehreren der folgenden Bereiche:

- Grundlagen relationaler Datenbanksysteme,
- Grundlagen von NoSQL-Datenbanksystemen,
- Datenmodellierung und Schemaentwurf,
- wissenschaftliches Arbeiten und Literaturrecherche,
- Programmiererfahrung, z. B. in Go, Python, Java oder einer vergleichbaren Sprache.

Vorkenntnisse im Bereich Benchmarking oder Datenbankmigration sind hilfreich, aber nicht zwingend erforderlich.

Startliteratur

- [1] Brian F. Cooper u. a. „Benchmarking cloud serving systems with YCSB“. In: *Proceedings of the 1st ACM Symposium on Cloud Computing, SoCC 2010, Indianapolis, Indiana, USA, June 10-11, 2010*. Hrsg. von Joseph M. Hellerstein, Surajit Chaudhuri und Mendel Rosenblum. ACM, 2010, S. 143–154. DOI: 10.1145/1807128.1807152. URL: <https://doi.org/10.1145/1807128.1807152>.
- [2] Asya Kamsky. „Adapting TPC-C Benchmark to Measure Performance of Multi-Document Transactions in MongoDB“. In: *Proc. VLDB Endow.* 12.12 (2019), S. 2254–2262. DOI: 10.14778/3352063.3352140. URL: <http://www.vldb.org/pvldb/vol12/p2254-kamsky.pdf>.

- [3] Jeyhun Karimov, Tilmann Rabl und Volker Markl. „PolyBench: The First Benchmark for Polystores“. In: *Performance Evaluation and Benchmarking for the Era of Artificial Intelligence - 10th TPC Technology Conference, TPCTC 2018, Rio de Janeiro, Brazil, August 27-31, 2018, Revised Selected Papers*. Hrsg. von Raghunath Nambiar und Meikel Poess. Bd. 11135. Lecture Notes in Computer Science. Springer, 2018, S. 24–41. DOI: 10.1007/978-3-030-11404-6_3. URL: https://doi.org/10.1007/978-3-030-11404-6%5C_3.
- [4] Shelly Sachdeva, Neha Bansal und Hardik Bansal. „SRank: Guiding schema selection in NoSQL document stores“. In: *Data Knowl. Eng.* 154 (2024), S. 102360. DOI: 10.1016/J.DATAK.2024.102360. URL: <https://doi.org/10.1016/j.datak.2024.102360>.