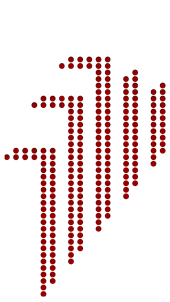

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Hochschule Darmstadt
- Fachbereich Informatik -

Benchmarking von NoSQL-Datenbanksystemen

Abschlussarbeit zur Erlangung des akademischen Grades
Master of Science (M.Sc.)

vorgelegt von
Holger Wegert

Referentin: Prof. Dr. Uta Störl
Korreferentin: Prof. Dr. Inge Schestag

Ausgabedatum: 01.10.2014
Abgabedatum: 01.04.2015

Zusammenfassung

NoSQL-Datenbanksysteme gewinnen im Kontext Big Data geprägter Problemstellungen kontinuierlich an Bedeutung. Einerseits steigt die Vielfalt speziell entwickelter Systeme, die eine dem Anwendungsfall gegenüber optimierte Einsatzfähigkeit ermöglicht. Andererseits entwickeln sich bereits etablierte NoSQL-Datenbanksysteme stetig weiter und erweitern mit optionalen Funktionalitäten das jeweilige Leistungsspektrum.

Eine für die Auswahl und den nachfolgenden Betrieb interessante Fragestellung adressiert die Auswirkungen auf das Leistungsvermögen infolge unterschiedlicher Systemkonfigurationen. Aus diesem Aspekt heraus wurde im Rahmen dieser Arbeit untersucht, inwiefern sich verschiedene Konfigurationsparameter bei den NoSQL-Datenbanksystemen Couchbase 3.0.1 und Apache Cassandra 2.1.2 auswirken. Der Verzicht auf eine direkte Gegenüberstellung der Datenbanksysteme ermöglicht eine den Systemen entsprechende Evaluierung, ohne dass einzelne Konfigurationsparameter gesetzt werden müssten, welche eine vermeintliche Vergleichbarkeit herstellen.

Für das Benchmarking der Systeme wurde Thumbtack Technologies Variante des Yahoo! Cloud Serving Benchmarks (YCSB) weiterentwickelt. Die durchgeführten Modifikationen umfassen zum einen funktionale Erweiterungen, wie bspw. den Ausbau der Mutli-Client Unterstützung. Zum anderen konnten bestehende Schwächen und Fehler identifiziert und behoben werden, welche andernfalls zu Verfälschungen der Messergebnisse führen.

Die Evaluierung der Datenbanksysteme hat gezeigt, dass die unterschiedlichen Konfigurationsmöglichkeiten eine Verwendung der Systeme bei verschiedenartigen Anforderungen grundsätzlich ermöglichen. Je nach Konfiguration wirkt sich deren Anpassung unterschiedlich stark auf das Leistungsvermögen der Systeme aus. Als besonders ausschlaggebend haben sich die unterschiedlichen Konfigurationsmöglichkeiten bzgl. der Dauerhaftigkeit und der Konsistenz herausgestellt. Beispielsweise führt bei beiden Datenbanksystemen ein garantiert dauerhaftes Schreiben zu einer Durchsatzverminderung von mehr als 97 %.

Abstract

NoSQL database systems gain in the context of Big Data affected problems continuously in importance. On the one hand increases the diversity of specially designed systems, which allow an optimized utilizability for each application. On the other hand, already established NoSQL database systems develop steadily and expand with optional functionalities their range of services.

One for the selection and subsequent operation interesting question addresses the effects of different system configurations on the performance of the system. Within the scope of this master thesis this aspect was taken into account and it was examined to what extend various configuration parameters in the NoSQL database systems Couchbase 3.0.1 and Apache Cassandra 2.1.2 impact. The forgoing of a direct comparison of the database systems allows an appropriate evaluation of the single systems without the definition of individual configuration parameters, which would establish an alleged comparability.

To establish an appropriate benchmark for the systems, Thumbtack Technologies version of Yahoo! Cloud Serving Benchmark (YCSB) was further developed. The realized modifications include on the one hand functional enhancements, such as the expansion of the multi-client support. On the other hand could existing weaknesses and errors be identified and corrected which otherwise would lead to mistakes in the measurement results.

The evaluation of the data base systems has shown that the different configuration possibilities in principle make the use of the systems at various requirements possible. Depending on the configuration its modulation affects to different extents the performance of the systems. It turned out that particularly decisive are different possibilities of configuration regarding the durability and consistency. For example, for both tested database systems an assured durable writing leads to a reduction in throughput of more than 97 %.