Praktikum 1: Datenmodellierung mit NoSQL-Datenbanksystemen

Ziel des Praktikums

Ziel des ersten Praktikums ist es, Erfahrungen mit der Datenmodellierung für dokumentenorientierte Datenbanksysteme – auch im Vergleich zu relationalen Datenbanksystemen – zu sammeln und sich mit den im Praktikum verwendeten Daten und Datenbanksystemen vertraut zu machen.

Hinweise:

- Grundlegende Informationen zur Nutzung der Datenbanksysteme finden Sie im Wiki: `https://wiki.h-da.de/fbi/bigdata/` → Software und Daten → Installierte Software bzw. direkt unter `https://wiki.h-da.de/fbi/bigdata/index.php/Installierte_Software` (Einloggen mit Ihrem hda-Account st*)
- Alle Praktikumsteilnehmer können sich auf den Rechnern des *Big Data Cluster Production* mit Ihrem LDAP-Account des Fachbereichs Informatik (ist*) einloggen. Für die Datenbanksysteme *MongoDB* und *Couchbase* bekommen Sie separate Accounts für ihre jeweiligen Gruppen (`prakN`).
- Den *PostgreSQL*-Server können Sie mit Ihrem LDAP-Account (ist*) benutzen. Rufen Sie eine ssh-Shell mit: istbenutzer@postgres.fbi.h-da.de auf. Ihre Datenbank wird dann automatisch angelegt und Sie können mit der Postgres-Shell oder mit PGAdmin arbeiten.
- Da diese Infrastruktur auch für andere Lehrveranstaltungen und studentische Abschlussarbeiten genutzt wird, ist es wichtig, dass Sie Ihre Daten sichern und Skripte so gestalten, dass ein schnelles Wiederaufsetzen Ihrer Datenbanken und Daten möglich ist.

Aufgabe 1 (Datenmodellierung - Teil 1)

Das folgende Szenario ist gegeben:

Die Forschungsgruppe **GroupLens** lässt Nutzer im Rahmen ihres Projektes **MovieLens** (`https://movielens.org`) Filme bewerten. Jeder Nutzer hat eine UserID und kann für jeden beliebigen Film eine Bewertung abgeben. Die Bewertungen liegen auf einer Skala von 1 bis 5 und sind in 0,5er Schritten möglich. Jeder Film hat einen Titel, in dem auch das Erscheinungsjahr enthalten ist, und ein oder mehrere Genres.

Modellieren Sie das gegebene Szenario zunächst als UML-Modell und bilden Sie es dann geeignet auf das relationale Datenmodell bzw. das dokumentenorientierte Datenmodell der verschiedenen Datenbanksysteme ab.

Dokumentieren Sie das UML-Modell und die spezifischen Datenmodelle geeignet in Ihrem Praktikumsbericht.

Begründen Sie an den Stellen, an denen Sie im Modell Wahlfreiheiten bei der Abbildung haben, Ihre getroffenen Entscheidungen. An Stellen, an denen Ihre Modellierungs- oder Abbildungsentscheidung vom Use Case abhängig ist, können Sie sich die Abfragen aus Praktikum 2 ansehen und diese als den Use Case annehmen.

Aufgabe 2 (Datenmodellierung - Teil 2)

Die originalen Datensätze mit Dokumentation finden Sie unter
`http://grouplens.org/datasets/movielens/`
bzw. bereits (schreibgeschützt) herunter geladen auf den Rechnern im Cluster unter
`/mnt/datasets/Movielens/original`

Vergleichen Sie jetzt Ihre Datenmodellierung mit der Modellierung die sich implizit aus den vorliegenden Daten ergibt. An Stellen an denen es Unterschiede zwischen Ihrer Modellierung und den Daten gibt, analysieren und bewerten Sie diese.

Aufgabe 3 (Einfügen von Daten)

Wählen Sie aus den Daten einige Beispiel-Datensätze aus, bringen diese in ein für das jeweilige Datenbanksystem geeignete Format entsprechend Ihres gewählten Datenmodells und fügen Sie diese in die jeweilige Datenbank ein. Dabei sollten mindestens fünf Filme und deren Bewertung eingefügt werden. Es sollen sowohl Filme mit einem als auch mit mehreren Genres enthalten sein und auch Filme ohne Bewertungen berücksichtigt werden.

Fügen Sie die Datensätze entweder einzeln ein oder (besser) wählen Sie eine Daten-Import-Variante, mit der Sie später (also in den folgenden Praktika) auch große Datenmengen importieren können. Hinweise dazu finden Sie im Wiki.

Alternative

Alternativ können Sie auch gleich die Massen-Import-Variante wählen und die 1M-Datensätze in die jeweilige Datenbank einfügen. Diese finden Sie aufbereitet in den Verzeichnissen. Wenn Sie diese Vorgehensweise wählen, müssen Sie mit dem sich aus den aufbereiteten Daten ergebendem Datenmodell weiterarbeiten oder die Daten anpassen.
- `/pgpool/movielens/adjusted` für PostgreSQL
- `/mnt/datasets/Movielens/JSON` für MongoDB und
- `/mnt/datasets/Movielens/couchbase` für Couchbase

Dokumentieren Sie im Praktikumsbericht kurz (d.h. nur exemplarisch das Grundprinzip), wie Sie die Daten in das jeweilige DBMS eingefügt haben.

Praktikumsbericht

Der Praktikumsbericht ist **eine Woche** nach dem Praktikumstermin per Mail als **pdf-Datei** an die jeweiligen Praktikumsbetreuer (also U. Störl für die Di-Gruppen und A. Hillenbrand und B. Reuschling für die Mo-Gruppe) abzugeben.