DATENBANKEN & SQL

– Organisatorisches & Einstieg –

Kurs am RRZK der Universität zu Köln

Rüdiger Voigt, M.A.

13.08.–17.08.2018

Übersicht

ORGANISATORISCHES

Rahmenbedingungen

- Termine: täglich vom 13.08. bis zum 17.08.2018, jeweils 10 bis 13 Uhr in diesem Raum
- *Es gibt keinen Abschlusstest.*
- *Sie erwerben keine Credit Points.*
- Wenn sie bei *allen* Terminen anwesend waren, erhalten Sie eine Teilnahmebestätigung.

Stil der Veranstaltung

Diese Veranstaltung ist kein Monolog!

Das bedeutet:

- Stellen Sie Fragen!
- Wir arbeiten gemeinsam auf einem Testsystem:
  - an jedem Kurstag kleine Übungen
  - ab Mittwoch ein Übungsprojekt in Kleingruppen

Pingo für Verständnisfragen und Feedback

Im Laufe der Veranstaltung werde ich gelegentlich Verständnisfragen stellen. Dazu nutzen wir das Pingo-System:

- Sie müssen keine Software installieren. Ein beliebiger Webbrowser reicht.
- Sie müssen sich nirgends registrieren. Sie benötigen nur die Session ID.
- *Ihre Antworten sind anonym.* Es ist mir technisch nicht möglich die Antwort einer konkreten Person zuzuordnen.

Folien

Die Folien finden Sie online unter

`https://www.ruediger-voigt.eu/kurs-datenbanken-und-sql.html`

Dort finden Sie ebenfalls *weiterführende Links und Literaturhinweise*.

Wichtige Informationen

Besonders wichtige Informationen sind auf den Folien *hervorgehoben* oder im **Fettdruck**.

Warnung / Hinweis

Den Inhalt solcher Boxen sollten Sie *auf keinen Fall ignorieren*. Hier finden Sie:

- Hinweise auf häufige Fehler, die zu Datenverlust führen können.
- Best Practices, die Ihre Anwendung um ein Vielfaches beschleunigen können.
- ...

Vorstellungsrunde

Stellen Sie sich *kurz* vor. Dabei interessieren:

- Ihr Name
- Ihre Fachrichtung
- Eventuelle Vorkenntnisse (Programmiererfahrung? Statistiksoftware? . . .)
- Haben Sie eine konkrete Anwendung für Datenbanken im Blick?
- Was erwarten Sie von diesem Kurs?

Ziele für diesen Kurs

- Sie lernen die Grundlagen im Umgang mit und der Konzeption von Datenbanken.
- Sie lernen die wichtigsten “Best Practices”.
- Sie vermeiden ab sofort die häufigsten Fehler.
- Sie stellen die richtigen Fragen und vertiefen Ihre Kenntnisse.
- Sie lernen Strategien zur Problemlösung und den zielgerichteten Gebrauch der Dokumentation.

EINFÜHRUNG

Was ist eine Datenbank?

Microsoft Excel ist keine Datenbank, sondern eine Software für Tabellenkalkulation.

Im engeren Sinne ist die Datenbank Ihre Zusammenstellung von Daten. Ein **Database Management System (DBMS)** verwaltet eine oder mehrere Datenbanken. Ihre Anwendung greift auf das DBMS zu. Die Formatierung / optische Darstellung ist nicht Aufgabe eines DBMS.

Umgangssprachlich wird oft der Begriff Datenbank anstelle von DBMS verwendet.

Zwecke / Fähigkeiten eines DBMS I

Die meisten relationalen DBMS verfügen über folgende Eigenschaften[1]:

- **dauerhafte und sichere Speicherung von Daten**
- **Mehrbenutzerbetrieb**
  Viele Personen arbeiten gleichzeitig mit dem selben Datensatz. Jedem Nutzer stehen stets die aktuellsten Daten zur Verfügung.
- **Transaktionen**
  Eine Kette von Aktionen wird etwa ganz oder gar nicht durchgeführt. (Zum Beispiel wird Ihr Konto nur belastet, wenn der Betrag einer Gegenstelle gutgeschrieben wird.)
- **Deduplikation**
  Vermeidung redundanter Datenhaltung und des mehrfachen Eingebens der selben Informationen.

Zwecke / Fähigkeiten eines DBMS II

- **Mandatenfähigkeit**
  Jeder Nutzer des Systems kann nur diejenigen Daten sehen / bearbeiten, welche für ihn freigegeben wurden.
- **relationale Systeme: Gewährleistung von Referentieller Integrität**
  Ein Datum kann nur dann gelöscht werden, wenn keine anderen Daten davon abhängen.
- **effizienter Zugriff auf (extrem) große Datenmengen und Teilmengen davon**
- **Grundlegende Analysen**
- **...**

[1] Es gibt Systeme, die aus gutem Grund andere Features haben.

Welche DBMS gibt es?

Es gibt sehr viele verschiedene DBMS für unterschiedliche Anwendungszwecke.

Im Kurs lernen Sie den Umgang mit relationalen Datenbanken. Das ist die dominante Technologie.

Es gibt daneben so genannte No-SQL-Datenbanken (Graph-Datenbanken, Key-Value Stores, ...) deren Bedeutung zunimmt. Ein Grund ist, dass sie oft sehr gut horizontal skalieren. Manchmal opfern Sie dafür aber wichtige Features, die relationale DBMS aufweisen. Diese Datenbanken verwenden oft Abwandlungen von SQL zur Bedienung.

Faustregel: Alles was mit Geld zu tun hat, läuft noch über relationale Datenbanken mit Transaktionen.

Was ist das beste DBMS?

- Es gibt nicht *das* beste DBMS. Jedes hat seine Stärken und Schwächen.
- Bei großen Projekten werden manchmal verschiedene DBMS parallel eingesetzt.
- Neben verfügbaren Features spielen oft die Lizenz und die Kosten eine entscheidende Rolle.

Das Kurs-Übungssystem

- Während des Kurses und für einige Wochen danach steht Ihnen online ein Übungssystem auf Basis von MariaDB zur Verfügung.
- Hier erfolgt kein Backup.
- Verwenden Sie es nicht für vertrauliche Daten.

Warum MariaDB? I

In diesem Kurs verwenden wir MariaDB. Ausschlaggebende Gründe für diese Wahl waren:

1. MariaDB ist ein Fork des sehr weit verbreiteten MySQL-DBMS und wurde als Drop-In-Replacement dafür konzipiert. Das bedeutet: Sie können Anleitungen und Problemlösungen für MySQL meist sehr leicht oder sogar 1:1 auf MariaDB übertragen.
2. Die Software ist ausgereift und wird aktiv weiterentwickelt. MariaDB wird aber schneller weiterentwickelt als MySQL.
3. MariaDB ist Free & Open Source Software (FOSS), während Sie für kommerzielle System leicht tausende Euro ausgeben können.
4. Das RRZK stellt Institutionen MariaDB Zugang zur Verfügung.

Warum MariaDB? II

5. MariaDB steht für viele Plattformen zur Verfügung. Es läuft unter anderem auf Linux, Windows und Mac. Sie können sich zum Lernen „mal eben“ das DBMS lokal installieren.

Alternative PostgreSQL

PostgreSQL ist wie MySQL/MariaDB Free & Open Source Software. Es ist ein mindestens gleichwertiges DBMS, aber nicht ganz so weit verbreitet.

Falls Sie sehr viel mit geographischen Daten arbeiten, dass ist PostgreSQL mit der PostGIS-Erweiterung wahrscheinlich eine deutlich bessere Wahl als MariaDB.

Dennoch sollten Sie während des Kurses MariaDB nutzen. Das Gelernte können Sie später leicht übertragen.

weitere wichtige relationale DBMS

kommerzielle Alternativen:

- **Microsoft SQL-Server**
  Wird oft in Verbindung mit SAP und ähnlichen Systemen verwendet. Die 2016er Variante integriert die Scriptsprache R. Es gibt eine Express-Version.
- **Oracle**
  Unter anderem im Finanzsektor weit verbreitet. Bekannt für teils sehr hohe Lizenzkosten.

SQL

- Die **Structured Query Language (SQL)** ist eine Programmiersprache, die bei fast allen relationalen Datenbanksystemen verwendet wird.
- Die Befehle ähneln der englischen Sprache.

SQL II

Hinweis

Der Kern der Sprache ist standardisiert, aber es gibt **Dialekte**. Das heißt:

- Bestimmte Konstrukte verhalten sich in anderen DBMS unerwartet oder funktionieren einfach nicht.
- Einge DBMS erweitern SQL um produktspezifische Funktionen.

Dialekte sind entstanden, weil einige DBMS älter sind als der Standard: nicht dem Standard entsprechende Befehle wurden beibehalten, damit Anwendungen weiterlaufen.
Manche DBMS erweitern SQL um neue Konstrukte und hoffen, dass diese in den Standard übernommen werden.

SQL III – verbreitete Dialekte

Einige verbreitete Dialekte:

- T-SQL / Transactional SQL: Microsoft SQL Server
- PL/pgSQL: PostgreSQL
- PL/SQL: Oracle Database

SQL ähnliche Sprachen

Falle

Einige NoSQL-DBMS verwenden Befehlssprachen, die SQL sehr ähnlich sind:

- Vorteil: Mit SQL-Kenntnissen fällt es Ihnen relativ leicht auch diese Systeme zu erlernen.
- Nachteil: Innerhalb von SQL können Sie sich trotz der Dialekte auf den Sprachkern / grundlegende Funktionen verlassen. Bei SQL ähnlichen Sprachen (zum Beispiel CQL im Cassandra DBMS) verhalten sich einige Befehle erheblich anders, wie sich das ein SQL Programmierer zusammenreimt.

GRUNDBEGRIFFE

Grundbegriffe I

Folgende Begriffe werden Sie immer wieder hören:

- **Datum**
  Singular von Daten
- **DBMS**
  Das Database Management System (DBMS) verwaltet Datenbanken.
- **Tabelle / Table**
  Eine relationale Datenbank besteht aus mehreren Tabellen, die in einem Zusammenhand stehen. Jede Tabelle hat Columns (Spalten) und Rows (Zeilen)

Grundbegriffe II

- **Server**
  Falls Ihre Datenbank nur lokal auf Ihrem Rechner läuft, können Sie nicht bzw. nur umständlich mit mehreren Personen gleichzeitig daran arbeiten. Ein Server ist ein Computer im Netzwerk, der dauerhaft in Betrieb ist. Wenn Ihre Datenbank dort betrieben wird, können mehrere Personen gleichzeitig darauf zugreifen.
  Als (HTTP-) Server bezeichnet man auch eine spezielle Software, welche auf einem Serversystem läuft.
- **GUI**
  Kurz für Graphical User Interface. Die Oberfläche eines Programms, das Sie mit der Maus bedienen.

Grundbegriffe III

- **Kommandozeile / Commandline / Commandline Interface (CLI)**
  Das Gegenteil eines GUI. Eine Eingabemaske, welche Ihre Befehle in Textform erwartet.
- **Query / Abfrage**
  Ein Befehl an das DBMS mit dem Sie Daten abfragen.
- **Function**
  Eine Sammlung von Befehlen, die Sie definieren können und die mit einem Aufruf abgearbeitet wird. Sie übergeben einen Wert und erhalten einen Wert zurück.

Grundbegriffe IV

- **Stored Procedure**
  Im Prinzip eine Funktion. Die genauen Unterschiede sind abhängig vom DBMS. Meist besteht der wesentliche Unterschied darin, dass Ihnen in Stored Procedures alle Befehle zur Verfügung stehen, in Functions aber nur ein Teil. Der Aufruf erfolgt auch anders.
- **Aggregate Function / Aggregatfunktion**
  Eine Reihe von eingebauten Funktionen, die jeweils auf die gesamte Spalte angewendet werden.
- **Trigger**
  Wenn Daten eingefügt, geändert oder gelöscht werden, wird – sofern ein entsprechender Trigger existiert – eine frei zu definierende Stored Procedure ausgelöst.

Grundbegriffe V

- **Index**
  Sie können für Columns einen Index anlegen. Das funktioniert ähnlich wie der Index in einem Buch. Es beschleunigt die Suche nach bestimmten Werten in fast allen Fällen *erheblich*.
- **Unique Identifier**
  Ein Merkmal welches nicht zweimal vorkommt und ein Datum daher eindeutig identifiziert. Wenn es keinen unique identifier gibt, legt man in der Regel eine Spalte mit einer ID-Nummer an.
- **Primary Key / Primärschlüssel**
  Ein Index auf einen Unique Identifier. Ist besonders wichtig, weil Sie oft nach dem eindeutigen Merkmal suchen werden beziehungsweise es nutzen werden um Tabellen zu verknüpfen.

DAS TESTSYSTEM

Testsystem

- Während des Kurses werden Sie SQL-Befehle und Konzepte mit einem MariaDB-System ausprobieren.
- Sie haben 24/7 Zugang zu diesem System. Am 30.09.2018 werden Ihre Zugänge deaktiviert und die von Ihnen angelegten Daten gelöscht.
- Individuelle Zugangsdaten erhalten Sie auf Papier.
- Verwenden Sie keine vertraulichen oder wichtigen Daten. (Der Administrator kann *Alles* sehen und es gibt kein Backup.)

phpMyAdmin

- phpMyAdmin ist eine sehr weit verbreitete Weboberfläche für MySQL-DBMS. Es macht keinen Unterschied, dass wir stattdessen MariaDB verwenden.
- Das GUI abstrahiert viel, gibt aber den generierten SQL-Code aus.
- phpMyAdmin ist Free and Open Source Software[2]
- Das ist bequem, *aber*:
  - *PHPMyAdmin muss auf einem Webserver wie Apache laufen*.
  - Wir werden uns auf SQL-Code konzentrieren, weil Sie diesen brauchen um komplexere Aufgaben zu erledigen.
  - Benutzen Sie einen Standardbrowser (Chrome, Firefox, Safari, NICHT Internet Explorer) in einer aktuellen Version.

[2] `https://www.phpmyadmin.net/`

MySQL Workbench

- Ein Programm mit allen Funktionen von phpMyAdmin. Darüber hinaus können Sie Datenbanken mit einer grafischen Oberfläche entwerfen und/oder abändern.
- Auf den Rechnern in diesem Raum vorinstalliert.
- Werden wir ab Mittwoch verwenden.

Editor

Notizen im Editor

Für Notizen verwenden Sie unbedingt einen Editor[a] und weder Microsoft Word noch Open Office Writer.
Word Processing Systeme (wie die beiden letztgenannten) ändern SQL-Befehle ab, so dass sie oft nicht mehr vom DBMS verstanden werden.

Ein guter Editor färbt SQL-Code auch ein, wenn er ihn erkennt. (In der Regel tut er das, wenn die Datei mit der Endung .sql abgespeichert wird. Text sollten Sie dann als Kommentar setzen.)

[a] notepad, notepad++, vim, emacs, ...

Übung 1

- Rufen Sie das Testsystem über die die phpMyAdmin Oberfläche auf und loggen Sie sich ein.
- Schauen Sie sich um:
  - Sie finden dort ein paar geteilte Datenbanken.
  - Ihre persönliche Test-Datenbank trägt als Namen Ihren Login.
- Falls Sie sich nicht einloggen können, klären wir das *jetzt*.