Modul Datenbanken

Einfuehrung

IFI Wintersemester 2016/17

Zur Person

Hintergrund

- Diplom Informatiker
- MSc Marine Microbiology
- PhD Bioinformatik an der Jacobs University
- Wissenschaftler (Bioinformatiker) am Max Planck Institute for Marine Microbiology, Bremen

Zur Person

Hintergrund

- Diplom Informatiker
- MSc Marine Microbiology
- PhD Bioinformatik an der Jacobs University
- Wissenschaftler (Bioinformatiker) am Max Planck Institute for Marine Microbiology, Bremen

Kontakt:

- linkedin:http://www.linkedin.com/in/renzokottmann
- twitter: @renzokott

Voraussetzungen

- Verstaendnis der Grundlagen der Informatik
- Kenntnisse
  - Programmierung
  - Java

Agile Developement

- Individuals and interactions over processes and tools
- Working software over comprehensive documentation
- Customer collaboration over contract negotiation
- Responding to change over following a plan

vom Agile Manifesto

Agile Learning

- Individuals and interactions over processes and tools
  - ⇒ Zusammen lernen *wichtiger* als strikte Teilnahmeregeln

Agile Learning

- Individuals and interactions over processes and tools
  - ⇒ Zusammen lernen *wichtiger* als strikte Teilnahmeregeln
- Working software over comprehensive documentation
  - ⇒ Gute Projektloesung *wichtiger* als gute Darstellung

Agile Learning

- Individuals and interactions over processes and tools
  - ⇒ Zusammen lernen *wichtiger* als strikte Teilnahmeregeln
- Working software over comprehensive documentation
  - ⇒ Gute Projektloesung *wichtiger* als gute Darstellung
- Customer collaboration over contract negotiation
  - ⇒ Zusammen arbeiten und lernen *wichtiger* als Beurteilung basierend auf Teilnahmeregeln

Agile Learning

- Individuals and interactions over processes and tools
  - ⇒ Zusammen lernen *wichtiger* als strikte Teilnahmeregeln
- Working software over comprehensive documentation
  - ⇒ Gute Projektloesung *wichtiger* als gute Darstellung
- Customer collaboration over contract negotiation
  - ⇒ Zusammen arbeiten und lernen *wichtiger* als Beurteilung basierend auf Teilnahmeregeln
- Responding to change over following a plan
  - ⇒ Ich pass mich **eher** Euch an statt Euch meinen Lehrplan aufzudruecken

Veranstaltungsplan und Projektarbeit

- Hier ist der Plan
  - Dezember im wesentlichen Vorlesungen
- Januar gemeinsame Projektarbeit
  - d.h. ich mache auch **mein** Projekt
  - iterative Erweiterung
  - anhand dessen weitere Konzepte erklaert werden
- So sieht die Projektarbeit aus

auf geht's :)

Hintergrund

- Daten werden immer wichtiger
  - Nicht nur bei dot.com Firmen wie:
    - Google, Facebook, Twitter etc.
- Daten sind vom allgemeinem öffentlichen Interesse
  - Begriffe wie 'Big Data', 'Metadata', 'Datenschutz' oder 'Abhörskandal' machen Presseschlagzeilen
- Daten wachsen enorm
  - Schaetzung: 1.2 Zettabyte im Jahr 2010
  - Alle 18 Monate verdoppelt
- Oekonomische Bedeutung
  - in Wirtschaft, Wissenschaft und Politik.
  - Stichworte 'Data Economy' und 'Data as a currency'

Was sind Daten...

- Es gibt mehrere Definitionen zu Daten (engl. data)
  - von denen nur einige frei zugänglich sind:

Definition: Merriam Webster

data noun plural but singular or plural in construction,often attributive\'dā-tə, 'da- also'dä-\

: facts or information used usually to calculate, analyze, or plan something

: information that is produced or stored by a computer

Definition: Merriam Webster

Full Definition of DATA

1: factual information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation

> < the data is plentiful and easily available—H. A. Gleason >
>
> < comprehensive data on economic growth have been published — N. H. Jacoby >

2: information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

3: information in numerical form that can be digitally transmitted or processed

Definition in der Wirtschaftsinformatik:

Gabler Wirtschaftlexikon Daten im Kontext der Wirschaftinformatik:

> Zum Zweck der Verarbeitung zusammengefasste Zeichen, die aufgrund bekannter oder unterstellter Abmachungen Informationen (d.h. Angaben über Sachverhalte und Vorgänge) darstellen.

Definition in der Wirtschaftsinformatik:

Gabler Wirtschaftlexikon Daten im Kontext der Wirschaftinformatik:

> Zum Zweck der Verarbeitung zusammengefasste Zeichen, die aufgrund bekannter oder unterstellter Abmachungen Informationen (d.h. Angaben über Sachverhalte und Vorgänge) darstellen.

Daraus kann man folgern, dass Daten

1. semantisch
   - Informationen sind, welche Fakten der realen Welt wiedergeben

2. syntaktisch
   - eine kodierte digitale Folge von Zeichen sind, die nur in einem bestimmten Bedeutungskontext zu einer richtigen Interpretation führen

138

138

- Die Zeichen bzw. Zahlenfolge "138" kann je nach Kontext
  - eine real existierende Hausnummer oder
  - eine Rechnungsnummer sein oder
  - ein geografischer Breitengrad (138 Grad West). darstellen

Kategorisierung von Daten

Man unterscheidet Datensammlungen häufig anhand des vorliegenden Strukturierungsgrades in:

- unstrukturierte Daten
  - Beispiele: Dokumente, beliebige Texte, Grafiken
- semistrukturierte Daten
  - fuehren einen Teil der Strukturinformationen mit sich
  - muessen keiner allgemeinen formalisierten Struktur entsprechen
    - Beispiele: Daten gespeichert in Extensible Markup Language (XML), JSON oder CSV
- strukturierte Daten
  - Muessen gemäß einem Datenmodell gleiche Struktur haben

... und was ist nun eine Datenbank?

Defintion Datenbank

ist eine zweckorientierte Sammlung von Daten. Daten werden dabei organisiert und strukturiert, um zweckmäßige Aspekte der Welt zu modellieren, sodass die Erfassung und Verarbeitung dieser Daten effizient unterstützt wird.

Analog und digital

- Analoge Sammlungen von Daten sind Datenbanken!
  - z.B. Aktenschränke oder Karteikartensammlungen
  - Historisch und konzeptionell Vorläufer von modernen elektronischen Datenbanken.
- Alle Datenbanken dienen der effizienten
  - Erfassung,
  - Speicherung
  - und Verarbeitung von Daten.

Nutzen von Datenbanken

- Häufigste Anwendung ist Beantwortung von Anfragen wie z.B.
  - welche Bücher von 'Douglas Adams' sind in dieser Bibliothek erhältlich
  - in welchem Regal befindet sich "Per Anhalter durch die Galaxis"?

Nutzen von Datenbanken

- Häufigste Anwendung ist Beantwortung von Anfragen wie z.B.
  - welche Bücher von 'Douglas Adams' sind in dieser Bibliothek erhältlich
  - in welchem Regal befindet sich "Per Anhalter durch die Galaxis"?

Das Erstellen einer neuen Datensammlung erfordert daher als ersten Schritt, Daten zu strukturieren.

Von unstrukturierten zu strukturierten Daten

- Ein neue Datensammlung beginnt häufig mit der Erfassung der Daten
  - durch Digitalisierung
  - oder liegen schon vor
    - z.B. in einfachen Textdokumenten wie Word-files, PDFs oder auch Internetseiten.

Erfassung

Daten der Modulteilnehmerinnen (Teilnehmerinnenliste)

Analog

Digital(isierung)

Wie?

Digital(isierung)

Wie?

Zur ersten Erfassung und Verwaltung bietet sich auch ein Tabellenkalkulationsprogramm wie z.B. Excel, OpenOffice oder online Tools wie Google Spreadsheets an.

Digital(isierung)

Wie?

Zur ersten Erfassung und Verwaltung bietet sich auch ein Tabellenkalkulationsprogramm wie z.B. Excel, OpenOffice oder online Tools wie Google Spreadsheets an.

Google Spreadsheet

Erkenntnisse??

Erkenntnis Vorschlag

- Als Datenbankstruktur ist eine Tabelle geeignet:
  - Ein Datensatz pro Zeile
  - Eine Eigenschaft (Attribut) pro Spalte
  - Erste Zeile enthält die Namen der Eigenschaften (anstatt eines Datensatzes)
  - Die Reihenfolge der Zeilen ist egal
  - Die Reihenfolge der Spalten ist egal

Daten Speicherung

- Files:
  - Nun befinden sich die Daten in einer Datei persistent im Dateisystem gespeichert, d.h. diese werden über die Laufzeit eines Programms oder des Computers hinaus existieren.
  - Nicht zwingend Excel-Format:
    - Excel ist ein binäres Format ist
  - In vielen Fällen reicht eine Textdatei
    - Z.B. das CSV-Format (Comma-separated values)

Comma Separated Values

- CSV-Dateien entsprechen Tabellen, gekennzeichnet durch:
  - jede Zeile durch ein Zeilenendezeichen
  - Spalten durch ein Trennsymbol wie
    - z.B. ein Komma ',' oder Semikolon ';'

Eine CSV-Datei kann man mit allen Textverarbeitungsprogrammen und auch Tabellenkalkulationsprogrammen bearbeiten werden.

Probleme dieser einfachen Struktur:

```
Viereck;Axel;26123;Oldenburg;
Huber;Ina;12345;FFM;0123/65235
Lustig;Olga;12345;Frankfurt;0123/45456
Mustermann;Erika;12345;Frankfurt;0123/45456
Henseler;Herwig;26197;Großenkneten;04435/388486 (Fax:388487)
Lustig;Peter;Frankfurt;0123/45456
Huber;Ina;3454;Dresden;0283/11111
Mustermann;Erika;;Bremen;436654
```

Probleme dieser einfachen Struktur:

```
Viereck;Axel;26123;Oldenburg;
Huber;Ina;12345;FFM;0123/65235
Lustig;Olga;12345;Frankfurt;0123/45456
Mustermann;Erika;12345;Frankfurt;0123/45456
Henseler;Herwig;26197;Großenkneten;04435/388486 (Fax:388487)
Lustig;Peter;Frankfurt;0123/45456
Huber;Ina;3454;Dresden;0283/11111
Mustermann;Erika;;Bremen;436654
```

- Mögliche Redundanzen
- Beziehungen werden nicht repräsentiert
- keine Festlegung von Datentypen
- unklare Eindeutigkeiten

Semistrukturierte Daten: XML

```xml
<?xmlversion="1.0"?>
<adressen>
    <adresse>
        <nachname>Lustig</nachname>
        <vorname>Peter</vorname>
        <plz>12345</plz>
        <ort>Frankfurt</ort>
        <telefon>0123/45456</telefon>
    </adresse>
    <!-- einige Eintraege ausgelassen -->
    <adresse>
        <telefon>436654</telefon>
        <nachname>Mustermann</nachname>
        <vorname>Erika</vorname>
        <ort>Bremen</ort>
    </adresse>
</adressen>
```

Semistrukturierte Daten: JSON

```
{
    "adressen": [{
        "nachname": "Lustig",
        "vorname": "Peter",
        "plz": "12345",
        "ort": "Frankfurt",
        "telefon": "0123/45456"
    },
    {
        "telefon": "436654",
                "nachname": "Mustermann",
        "vorname": "Erika",
        "ort": "Bremen"

    }]
}
```

Semistrukturierte Daten

- Für einfache Anwendungen kann diese Form der Datenspeicherung und verwaltung in Datei(en) durchaus ausreichen.
- Filesysteme unterstuetzen nicht (oder nur unzureichend):
  - effiziente Suche und Modifikation von kleinen Dateneinheiten,
  - komplexe Datenanfragen,
  - Transaktionen
  - effizientes buffering und caching von Daten im Hauptspeicher

Anforderungen an Datenbanken in der Regel deutlich höher.

Danke fuer die Zusammenarbeit