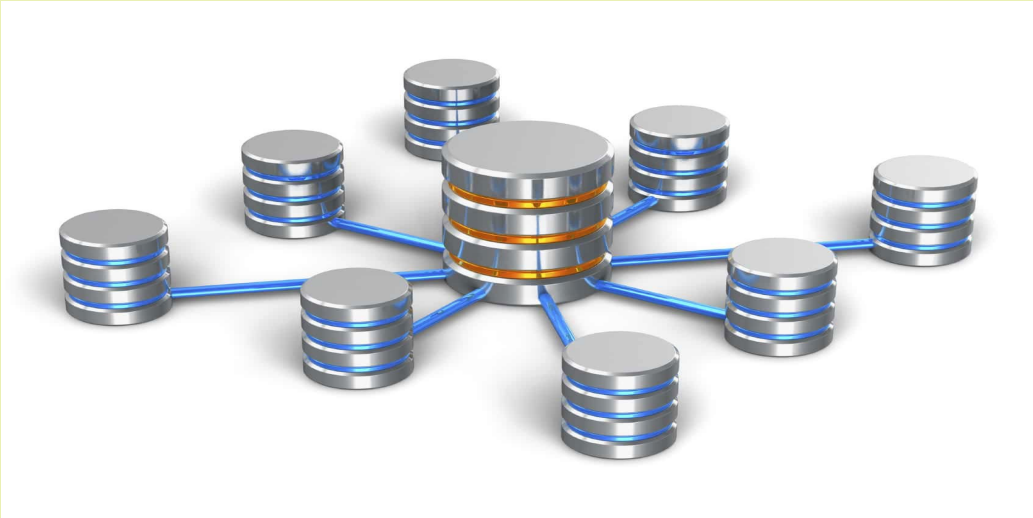
FRANKFURT UNIVERSITY OF APPLIED SCIENCES

3: Herausforderungen der Datenintegration

Prof. Dr. Markus Grüne, FB03
Wirtschaftsinformatik



Lernziele

- Verstehen, wie der Idealzustand einer Integration wäre → "ideal integration state"
- Warum sind Distribution, Autonomie und Heterogenität hinderliche bei der Integration?
- Die Gründe für Heterogenität nachvollziehen können.
- Ausprobieren: Übersetzung eines Schemas in das JSON-Format.

Das ideale Niveau der Integration erreichen

Für verlässliche und sichere Netzwerke sorgen.

- z.B. Verschlüsselung der Verbindungen und Schnittstellen

Vertauenswürdige Kommunikationskanäle aufsetzen → Agreements zwischen Endpunkten (SLA)

- Agreements spezifizieren die Nachrichten, die ausgetauscht werden, insbes.
  - deren Definition, Richtung
  - die Austauschsequenz
  - die Dauer, für die das Austausch-Agreement gültig ist
  - Und viele weitere Attribute
  - Auch: non-repudiation in B2B

(Bussler 2003)

Das ideale Niveau der Integration erreichen

Überbrückung semantischer Unterschiede

- Einheitliches semantisches Datenmodell und Integrationsverhalten
- Z.B. nur eine Definition für "purchase order" und alle anderen Prozesse

➢Nur dann kann eine homogene Integration erreicht werden.

Leider ist die Welt nicht ideal ☹

Herausforderungen - Kategorien

- Distribution
- Autonomie
- Heterogenität

Die folgenden Folien orientieren sich an (Leser, Naumann 2007)

Distribution / Verteilung

Verteilung in zwei Kategorien

- **Physisch**: Daten auf physisch separaten Systemen
- **Logisch**: z.B. wenn ein Datensatz auf mehreren physischen Systemen verteilt

<u>Probleme / Ansätze der physischen Verteilung</u>

Netzwerkebene: Physische Netzwerkknoten müssen identifizierbar und ansprechbar sein (Server, Port) → TCP/IP.

Daten oft nur in unterschiedlichen Schemas auf den Netzwerkknoten verteilt.

Es werden mehrere Query Languages für Retrieval benötigt oder eine, die mehrere Schemas beherrscht.

Abfrageoptimierung sollte die Netzwerklast berücksichtigen.

Distribution / Verteilung

Probleme der Logischen Verteilung:

- Überlappende Daten mit demselben Inhalt und derselben Bedeutung.
- Redundanzen
- Lokalisierungsprobleme: User sind nicht in der Lage, den korrekten Orten der Daten herauszufinden → IIS benötigt
- Duplikate
- Widersprüchliche Daten

Autonomie

Datenquellen (Teams) entscheiden selbstständig darüber, wie die Datenstrukturen aussehen und wie der Zugriff auf sie erfolg.

- **Interface** Autonomie
- **Design** Autonomie
- **Access** Autonomie
- **Rechtliche** Autonomie

Design Autonomie

Designer der Datenquellen entscheiden unabhängig über:

- Data format
- Data model
- Schema
- Syntax
- Keys
- Werteeinheiten ...

Interface, Access, rechtliche Autonomie

Datenquellen entscheiden über technische Aspekte

- Protokolle
- Abfragesprachen

Datenquellen entscheiden über den Zugriff einzelner Personen oder Rollen

- Userkonten
- Authentifizierung
- Autorisierung

Datenquellen können die Integration ihrer Daten ablehnen

- Copyrights

Arten der Heterogenität

- Technische Heterogenität: Probleme der technischen Umsetzung von Zugängen
- Syntaktische Heterogenität: Darstellung der Information
- Data Model Heterogenität / Strukturelle Heterogenität
- Schematische Heterogenität : Unterschiede im Datenschema
- Semantische Heterogenität : Unterschiede hinsichtlich der Bedeutung

Achtung: die Aspekte sind nicht völlig trennscharf

Beispiele für technische Heterogenität

Level	**Variants**
Query Facility	query language, parametrized functions, input forms
Query Language	SQL, XQUERY, full text search
Austauschformat	binary data, XML, HTML, tabular
Kommunikationsprotokoll	HTTP, JDBC, SOAP

(Leser, Naumann 2003)

Syntaktische Heterogenität

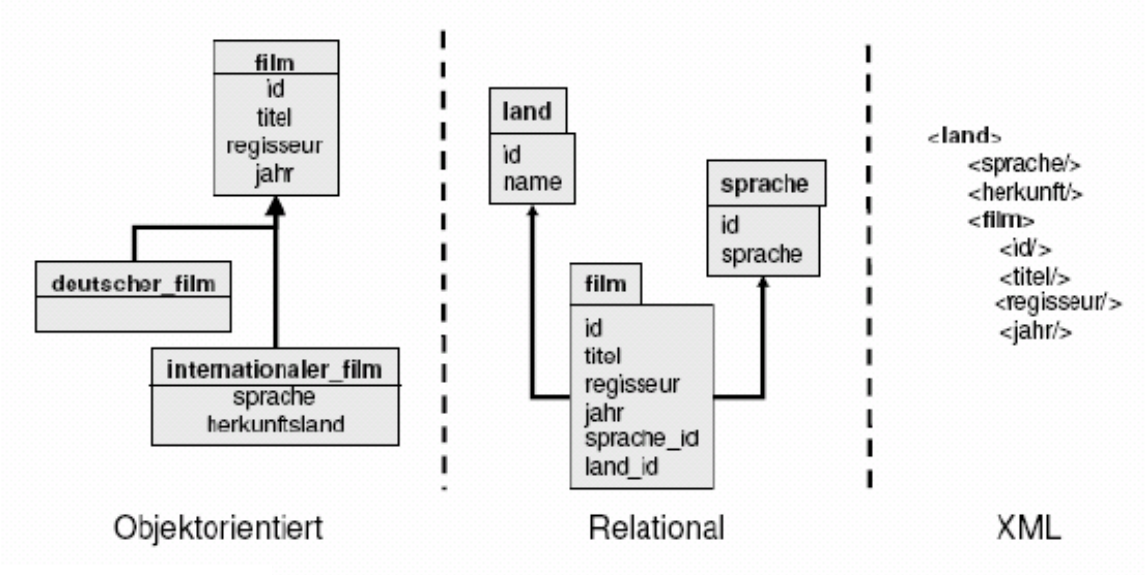
Unterschiedliche Repräsentation desselben Fakts

- Locales: Dezimalzeichen: Punkt oder Komma
- Euro oder €
- Comma-separated oder tab-separated
- EBCDIC oder ASCII oder Unicode
- Noten: A – F, „sehr gut“, „gut“, …
- Binary coding oder characters
- Datenformate (12. September 2016; Sept 12, 2016; 12.9.2016; 9/12/2016; 16092016…)

Straightforward Lösung

- Transformation durch Berechnung oder Übersetzungstabellen

Heterogenes Datenmodell



Objektorientiert

(Leser, Naumann 2003)

NoSQL - Beispiel

A JSON Document, as stored e.g. in MongoDB

```
{ "_id" : 1,
"name" : { "first" : "John", "last" : "Backus" },
"contribs" : [ "Fortran", "ALGOL", "Backus-Naur Form", "FP" ],
"awards" : [
{ "award" : "W.W. McDowell Award", "year" : 1967, "by" :
"IEEE Computer Society" },
{ "award" : "Draper Prize", "year" : 1993,
"by" : "National Academy of Engineering" } ]
}
```

[http://www.mongodb.com/json-and-bson]
Arrays in [...]
Subdocuments in { ... }
→ no 1st Normal Form

Übung

Erzeugen Sie ein JSON Document für den Kunden Shipston zusammen mit seinen Orders

CUSTOMER_ID	LASTNAME	POSTCODE	PLACE
100	Brown	TR197LU	St Just
101	Ryan	EC2M2RB	London
102	Smith	NE45AU	Corbridge
103	Shipston	W45LL	London

ORDER_ID	DELIVERYDATE	CUSTOMER_ID
151	02.05.2012	101
152	02.05.2012	103
153		105
154		103

Hinweis:
Arrays in […]
Subdocuments in { … }

Strukturelle und semantische Heterogenität

Strukturelle

- Zwei Schemas, die dasselbe Objekt beschreiben, sind unterschiedlich

Semantisch

- Elemente für zwei Schemas überlappen sich teilweise, d.h. teilweise wird dasselbe Objekt gespeichert mit unterschiedlicher Bedeutung / Verwendung.

(Leser, Naumann 2003)

Gründe für Heterogenität

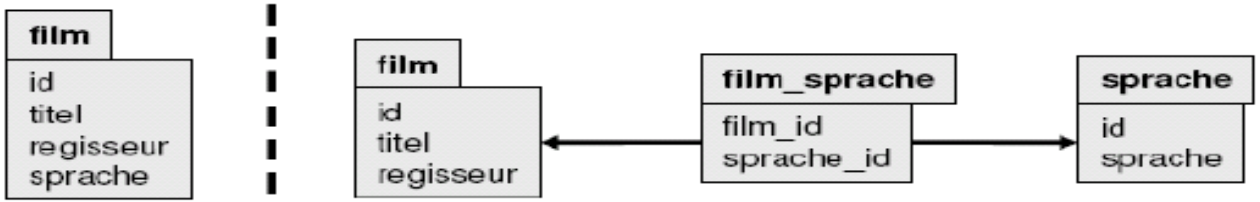
- Software Engineer hat eigene Präferenzen
- unabhängige, unkoordinierte Entwicklung
- Technische Normen
- Individuelle Anpassungen von Software
- "Wahlfreiheit" hinsichtlich des logischen Datenmodells wenn ein konzeptuelles Datenmodell gegeben ist (ERM oder Klassendiagramm)

Beispiel: Probleme mit Kardinalitäten

Film hat eine Sprache versus Film hat mehrere Sprachen.

Was ist ein Film?

Die Bedeutung des Wortes ist abhängig vom Kontext



Schemaheterogenität

Modellierung als Relationen

```
spielfilm      ( id, titel, laenge)
dokumentarfilm( id, titel, laenge)
```

Modellierung als Attribute

```
film( id, titel, laenge, spielfilm, doku)
```

Modellierung als Attributwerte

```
film( id, titel, laenge, typ)
```

Semantische Heterogenität

Was ist die Bedeutung eines bestimmten Tabellennamens?

Was ist die Bedeutung eines bestimmten Attributnamens?

Synonyme

Homonyme

Muss vermieden werden!!!

Key Takeaways

- Ein idealer Zustand der Integration wird selten erreicht und ist dann nicht von Dauer.
- Datenformate sind divers. Bei der Integration sind u.a. syntaktische und semantische Eigenheiten der Formate zu beachten.
- Herausforderungen der Integration bestehen in der
  - Distribution der Daten
  - Autonomie / den “Eigentumsverhältnissen”
  - Heterogenität

Literatur

Bussler C (2003) B2B integration; Concepts and architecture ; with 4 tables. Springer, Berlin

Leser U, Naumann F (2007) Informationsintegration; Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. dpunkt-Verl., Heidelberg