



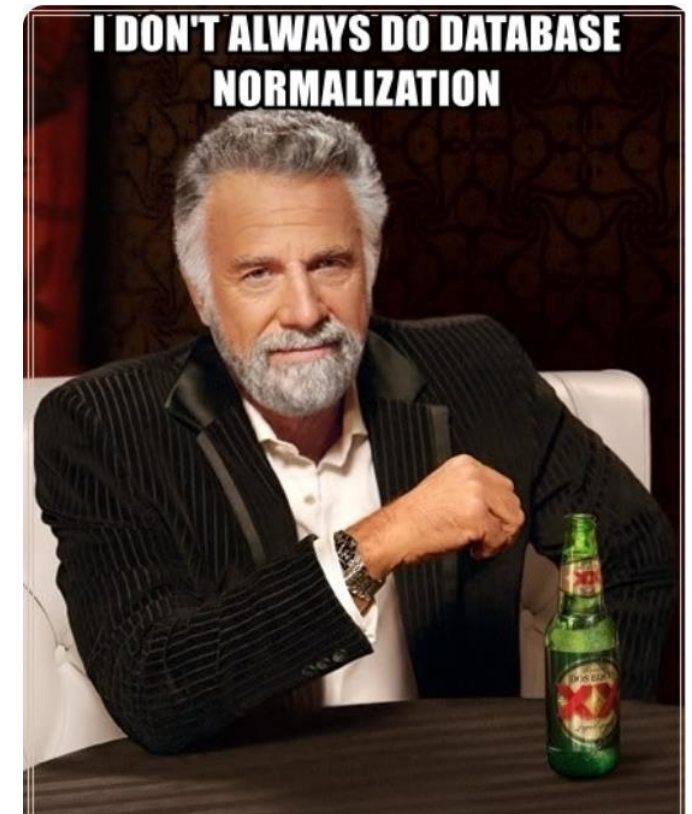
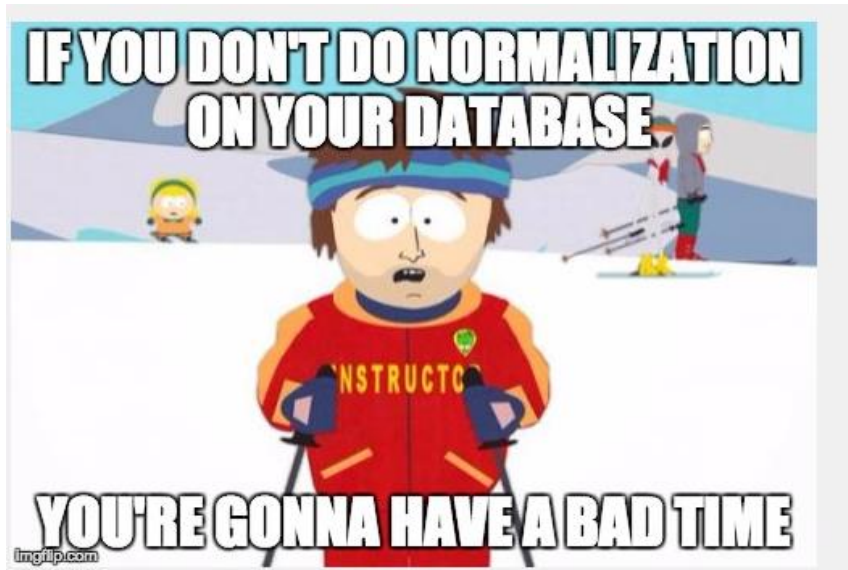
FFI / Datenbanksysteme

Normalisierung

Prof. Dr. Patrick Cato

Technische Hochschule
Ingolstadt





Normalisierung

- Aufteilung von Attributen in mehrere Relationen
- Redundanz von Daten ist schwer zu verwalten (duplication effort)
- **Grundannahme:** Möglichst **redundanzfreie Speicherung** von Daten, um **Anomalien** zu verhindern. Fremdschlüsselredundanzen werden akzeptiert.

Denormalisierung

- Übersetzung des Entity Relationship Modells in ein relationales Schema mit möglichst wenig Relationen (Zusammenfassen wo möglich)
- Aus Effizienzgründen ist sinnvoll, Daten mehrfach zu halten. Dies wird streng überwacht (controlled redundancy)
- **Grundannahme:** Der **JOIN** ist eine sehr **ressourcenintensive Operation**. Zur **Performanceoptimierung** kann es gerechtfertigt sein, Daten bewusst redundant zu halten und Anomalien zu akzeptieren.

Theoretische Grundlagen



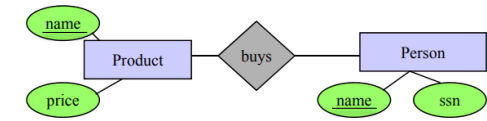
Definition Normalisierung

Als Normalisierung bezeichnet man die verlustfreie Zerlegung (nonloss / lossless decomposition) von Relationen. Hierbei gilt der Grundsatz: keine Information darf verloren gehen.

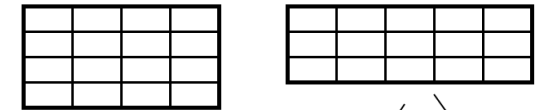
Durch Bildung eines JOINS müssen die zerlegten Informationen wiederherstellbar sein.

Relational Schema Design

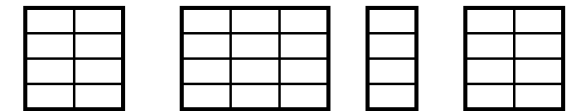
Conceptual Model:



Relational Model:
plus FD's



Normalization:
Eliminates anomalies



Was sind Probleme?

<u>InvNr</u>	ISBN	Title	Specialization	Author
2049	3-8538	Databases	Information System	{Saake, Sattler, Heuer}
2050	3-8538	Databases	Information System	{Saake, Sattler, Heuer}
3587	3-6633	Data Science	Information System	{Grunert, Meyer, Heuer}
4812	3-6633	Data Science	Information System	{Grunert, Meyer, Heuer}
4961	3-1007	Machine Learning	Artificial Intelligence	{Koste, Korbmacher}

Definition

Eine Einfügeanomalie ist eine Art von Dateninkonsistenz, die in einer Datenbank auftritt, wenn das Hinzufügen eines neuen Datensatzes zu einem Verstoß gegen Datenintegritätsregeln oder der Einführung redundanter Daten führt.

<u>InvNr</u>	<u>ISBN</u>	<u>Title</u>	<u>Specializatzion</u>	<u>Author</u>
4961	3-8538	Machine Learning	Artificial Intelligence	{Koste, Korbmacher}

<u>InvNr</u>	<u>ISBN</u>	<u>Title</u>	<u>Specialization</u>	<u>Author</u>
2049	3-8538	Databases	Information System	{Saake, Sattler, Heuer}
2050	3-8538	Databases	Information System	{Saake, Sattler, Heuer}
3587	3-6633	Data Science	Information System	{Grunert, Meyer, Heuer}
4812	3-6633	Data Science	Information System	{Grunert, Meyer, Heuer}
4961	3-1007	Machine Learning	Artificial Intelligence	{Koste, Korbmacher}

Definition

Eine Änderungsanomalie ist eine Art von Dateninkonsistenz, die in einer Datenbank auftritt, wenn Änderungen an bestehenden Informationen nicht korrekt oder konsistent im gesamten Datensatz übertragen werden wenn nicht alle (redundanten) Vorkommen eines Attributwertes zugleich geändert werden. Dieses führt zu inkonsistenten Daten.

Beispiel: Bücher mit dem Titel „Databases“ werden jetzt dem Themengebiet Big Data zugeordnet.

<u>InvNr</u>	ISBN	Title	Specialization	Author
2049	3-8538	Databases	Information System	{Saake, Sattler, Heuer}
2050	3-8538	Databases	Information System	{Saake, Sattler, Heuer}
3587	3-6633	Data Science	Information System	{Grunert, Meyer, Heuer}
4812	3-6633	Data Science	Information System	{Grunert, Meyer, Heuer}
4961	3-1007	Machine Learning	Artificial Intelligence	{Koste, Korbmacher}

Definition

Eine Löschanomalie ist eine Art von Dateninkonsistenz, die in einer Datenbank auftritt, wenn das Löschen eines Datensatzes unbeabsichtigt zum Verlust anderer, nicht zusammenhängender Daten führt.

Beispiel: Lösche Exemplare 4961 (Verlust)

<u>InvNr</u>	ISBN	Title	Specialization	Author
2049	3-8538	Databases	Information System	{Saake, Sattler, Heuer}
2050	3-8538	Databases	Information System	{Saake, Sattler, Heuer}
3587	3-6633	Data Science	Information System	{Grunert, Meyer, Heuer}
4812	3-6633	Data Science	Information System	{Grunert, Meyer, Heuer}
4961	3-1007	Machine Learning	Artificial Intelligence	{Koste, Korbmacher}

Informationsverlust



Definition

Gegeben sind zwei Attributmengen X und Y mit $X, Y \subseteq [R]$.

$X \rightarrow Y$ heißt funktionale Abhängigkeit (Functional Dependency, FD), falls gilt:

Es existiert eine Funktion

$f(x) := y$

für $x \in \pi_X(R)$, $y \in \pi_Y(R)$ für **alle möglichen Instanzen** von R .

**„Die Werte von x bestimmen die Werte von y eindeutig“
 X heißt Determinante.**

<u>persnr</u>	name	vorname	geburtstag	<u>projektnr</u>	pname	prioritaet
1	Schweitzer	Albert	01.03.1973	5	Unis	7
2	Carlos	Rob	12.07.1975	1	Data Center	10
2	Carlos	Rob	13.07.1975	3	Lobbysiet	8
2	Carlos	Rob	14.07.1975	6	Rabbit breeder	2
3	Mueller	Peter	09.10.1963	2	House breeder	3
3	Mueller	Peter	09.10.1963	4	Politician	5

$\{\text{persnr}\} \rightarrow \{\text{name, vorname, geburtsdatum}\}$

$\{\text{projektnr}\} \rightarrow \{\text{pname, prioritaet}\}$

$\{\text{projektnr}\} \rightarrow \{\text{persnr, name, vorname, geburtsdatum}\}$

...



Instanz versus Schema

Die funktionale Abhängigkeit bezieht sich auf das Schema und nicht die vorliegende Instanz (eine Tabelle mit Zeilen und Spalten bzw. der aktuelle Inhalt der Relation). FD definiert den Constraint für die **möglichen Instanzen** von R. Datenbankdesigner müssen diese FDs identifizieren.

Personen			
id <u>integer</u>	name character varying	vorname character varying	geburtsdatum date
1	Schweitzer	Albert	1973-03-01
2	Carlos	Rob	1975-07-12
3	Mueller	Peter	1963-10-09
4	Zappa	Frank	1955-11-04
5	Taylor	Tim	1980-03-04
6	Wurst	Hans	1974-02-01
7	Miese	Peter	1983-05-06
8	Koenig	Dieter	1967-06-11

$\{id\} \rightarrow \{name, vorname, geburtsdatum\}$ // ja
 $\{name\} \rightarrow \{vorname\}$ // nein
 $\{geburtsdatum\} \rightarrow \{name, id, vorname\}$ // nein

Zwar wären die FDs für die vorliegende Instanz korrekt, aber nicht für das Schema!

3.4 Geben ist das Relationenschema:

Angebot (Kaffee, Bohne, Land, Röstung, Verarbeitung, Typ)

Unten ist eine Ausprägung/Instanz (nach Kaffee sortiert) abgebildet.

Angebot					
Kaffee	Bohne	Land	Röstung	Verarbeitung	Typ
Alice	Arabica	Brasilien	plus	natur	rein
Bob	Arabica	Ecuador	plus	nass	rein
Carol	Arabica	Peru	full	nass	mischung
Carol	Arabica	Honduras	full	natur	mischung
Carol	Robusta	Indien	plus	nass	mischung
Dan	Arabica	Brasilien	full	nass	mischung
Dan	Robusta	Indien	full	natur	mischung
Eve	Arabica	Indonesien	full	nass	rein
Faythe	Arabica	Äthiopien	plus	natur	mischung
Faythe	Arabica	Guatemala	plus	nass	mischung
Grace	Arabica	Äthiopien	normal	nass	rein
Rupert	Robusta	Indien	full	nass	rein

Überprüfen Sie für jede der unten angegeben funktionalen Abhängigkeiten (FD), ob sie auf der gegebenen Ausprägung gelten oder nicht. Geben Sie für jede FD die Antwort (ja/nein) an. Falls eine FD nicht erfüllt ist geben Sie außerdem ein entsprechendes Beispiel als Begründung an.

Land → Bohne

a) Land → Bohne

b) Bohne → Land

c) (Kaffee, Bohne) → Land

d) Land → (Bohne, Verarbeitung)

Definition

Eine funktionale Beziehung, deren Determinanten irreduzibel ist, heißt volle funktionale Beziehung.

α und β sind Attributmengen eines relationalen Schemas $\text{sch}(R)$. Eine volle funktionale Abhängigkeit besteht wenn:

- Die funktionale Abhängigkeit $\alpha \rightarrow \beta$ gilt
- die Attributmenge α nicht verkleinert werden kann

In anderen Worten: Eine vollständig funktionale Abhängigkeit liegt dann vor, wenn das Nicht-Schlüsselattribut nicht nur von einem Teil der Attribute eines zusammengesetzten Schlüsselkandidaten funktional abhängig ist, sondern von allen Teilen.

Beispiel volle funktionale Abhängigkeit



<u>Reihe</u>	<u>Band</u>	Titel
Asterix	1	Asterix der Gallier
Asterix	17	Die Trabantenstadt
Asterix	25	Der große Graben
Tim and Struppi	1	Der geheimnisvolle Stern
Franka	1	Das Kriminalmuseum
Franka	2	Das Meisterwerk

Definition

Eine partielle Abhängigkeit ist eine Art von funktionaler Abhängigkeit, die in einer Relation zwischen Attributen existiert, wenn ein Nichtschlüsselattribut nur von einem Teil (Untermenge) eines zusammengesetzten Schlüssels (bestehend aus mehreren Attributen) abhängt, anstatt vom gesamten Schlüssel.

Definition

- Geht man von einer Attributmenge α aus $\text{sch}(R)$ aus, die ein Schlüssel ist und die Attributmenge β aus $\text{sch}(R)$ funktional bestimmt
- Dann liegt eine transitive Abhängigkeit vor, wenn β auch eine weitere Attributmenge γ aus $\text{sch}(R)$, die nicht Teil des Schlüssels ist, bestimmt. Also $\alpha \rightarrow \beta \rightarrow \gamma$

Buchexemplar			
<u>InvNr</u>	ISBN	Titel	Fachgebiet
2049	3-8538	Datenbanken	Informationssysteme
2050	3-8538	Datenbanken	Informationssysteme
2051	3-8538	Datenbanken	Informationssysteme
2121	3-4711	Formale Sprachen	Theoretische Informatik
3587	3-6633	Data Science	Informationssysteme
4812	3-6633	Data Science	Informationssysteme
4961	3-1007	Maschinelles Lernen	Künstliche Intelligenz

Normalformen

Definition

Eine Relation ist in erster Normalform (1NF), wenn alle Attribute **atomare Wertebereiche** haben (d.h. keine zusammengesetzten Wertebereiche)

Buchexemplar				
<u>InvNr</u>	ISBN	Titel	Fachgebiet	Autoren
2049	3-8538	Datenbanken	Informationssysteme	{ Saake, Sattler, Heuer }
4812	3-6633	Data Science	Informationssysteme	{ Grunert, Meyer, Heuer }
4961	3-1007	Maschinelles Lernen	Künstliche Intelligenz	{ Korste, Korbmacher }



- Die Relation Buchexemplar lässt sich in die erste Normalform überführen, in dem für jeden Eintrag in Autoren ein Tupel gebildet wird
- Damit entstehen allerdings weitere Redundanzen, die mit den folgenden Normalformen eliminiert werden müssen

Buchexemplar				
<u>InvNr</u>	<u>ISBN</u>	<u>Titel</u>	<u>Fachgebiet</u>	<u>Autoren</u>
2049	3-8538	Datenbanken	Informationssysteme	Saake
2049	3-8538	Datenbanken	Informationssysteme	Sattler
2049	3-8538	Datenbanken	Informationssysteme	Heuer
4812	3-6633	Data Science	Informationssysteme	Grunert
4812	3-6633	Data Science	Informationssysteme	Meyer
4812	3-6633	Data Science	Informationssysteme	Heuer
4961	3-1007	Maschinelles Lernen	Künstliche Intelligenz	Korste
4961	3-1007	Maschinelles Lernen	Künstliche Intelligenz	Korbmacher

Ab wann ein Wert als atomar angesehen wird, hängt vom Nutzungskontext ab. Beispielsweise ist die Trennung von mehreren Vornamen in einzelne Spalten nicht sinnvoll, weil es keine Abfragen auf die einzelnen Vornamen gibt. Zudem gibt es in manchen Kulturkreisen viele zusammengesetzte Namen, sodass auch nicht bekannt wäre, wie viele Spalten anzulegen wären. Es ist daher immer auch auf die Machbarkeit zu achten!

Definition

Eine Relation ist in der 2. Normalform (2NF) wenn:

- Eine Relation in der 1. Normalform ist
- Wenn jedes Nichtschlüsselattribut von **allen Schlüsselkandidaten** voll (irreduzierbar) funktional abhängt.

Oder anders: Eine Relation R liegt nicht in der zweiten Normalform vor, wenn es ein Nichtschlüsselattribut gibt, das nur von einem Teil des Schlüssels abhängt.

Die Relation Buch ist nicht in 2NF

Buch			
<u>ISBN</u>	<u>Titel</u>	<u>Fachgebiet</u>	<u>Autoren</u>
3-8538	Datenbanken	Informationssysteme	Saake
3-8538	Datenbanken	Informationssysteme	Sattler
3-8538	Datenbanken	Informationssysteme	Heuer
3-6633	Data Science	Informationssysteme	Grunert
3-6633	Data Science	Informationssysteme	Meyer
3-6633	Data Science	Informationssysteme	Heuer
3-1007	Maschinelles Lernen	Künstliche Intelligenz	Korste
3-1007	Maschinelles Lernen	Künstliche Intelligenz	Korbmacher

Jede Relation R , die nicht in 2NF ist, wird folgendermaßen zerlegt:

Die Relation $R[A, B, C]$ (auf den schnittfreien Attributmengen A, B, C) habe die irreduzierbare FD: $A \rightarrow B$, wobei A echter Teil eines Schlüssels ist und B ein Nichtschlüsselattribut ist, dann wird durch die Zerlegung gebildet:

$R_1 = R[\underline{A}, B]$

$R_2 = R[\uparrow \underline{A}, C]$ oder $R[\uparrow \underline{A}, \underline{C}]$

A ist Fremdschlüssel in R_2

Buch			
<u>ISBN</u>	<u>Titel</u>	<u>Fachgebiet</u>	<u>Autoren</u>
3-8538	Datenbanken	Informationssysteme	Saake
3-8538	Datenbanken	Informationssysteme	Sattler
3-8538	Datenbanken	Informationssysteme	Heuer
3-6633	Data Science	Informationssysteme	Grunert
3-6633	Data Science	Informationssysteme	Meyer
3-6633	Data Science	Informationssysteme	Heuer
3-1007	Maschinelles Lernen	Künstliche Intelligenz	Korste
3-1007	Maschinelles Lernen	Künstliche Intelligenz	Korbmacher

Relation Buchexemplar lässt sich **durch Zerlegung** in zwei Relationen in die zweite Normalform bringen

Buch			Buch_Autoren	
<u>ISBN</u>	<u>Titel</u>	<u>Fachgebiet</u>	<u>ISBN</u>	<u>Autor</u>
3-8538	Datenbanken	Informationssysteme	3-8538	Saake
3-6633	Data Science	Informationssysteme	3-8538	Sattler
3-1007	Maschinelles Lernen	Künstliche Intelligenz	3-8538	Heuer
			3-6633	Grunert
			3-6633	Meyer
			3-6633	Heuer
			3-1007	Korste
			3-1007	Korbmacher



Bilden Sie eine 3er oder
2er Gruppe und bearbeiten
Sie das Aufgabenblatt

Gesamtzeit: 15 Minuten

Aufgabe 1



<u>Bestellnummer</u>	<u>Artikelnummer</u>	Artikelname	Bestelldatum	Kundennummer	Kundenname
1	100	Laptop	2023-04-01	123	Alice
1	200	Maus	2023-04-01	123	Alice
2	100	Laptop	2023-04-05	456	Bob
3	300	Tastatur	2023-04-10	123	Alice

Gegeben sind die vollen funktionalen FDs:

Bestellnummer → Bestelldatum

Bestellnummer → Kundennummer

Artikelnummer → Artikelname

Kundennummer → Kundenname

...

Schlüsselkandidaten:

{Bestellnummer, Artikelnummer}

- **Warum ist die Relation nicht in 2NF?**
- **Überführen Sie durch Zerlegung in die 2. Normalform**

Buch			
ISBN	Titel	Fachgebiet	Autoren
3-8538	Datenbanken	Informationssysteme	Saake
3-8538	Datenbanken	Informationssysteme	Sattler
3-8538	Datenbanken	Informationssysteme	Heuer
3-6633	Data Science	Informationssysteme	Grunert
3-6633	Data Science	Informationssysteme	Meyer
3-6633	Data Science	Informationssysteme	Heuer
3-1007	Maschinelles Lernen	Künstliche Intelligenz	Korste
3-1007	Maschinelles Lernen	Künstliche Intelligenz	Korbmacher



In welcher NF befindet sich die untenstehende Tabelle?

<u>id</u>	ISBN	Titel	Fachgebiet	Autoren
1	3-8538	Datenbanken	Informationssysteme	Saake
2	3-8538	Datenbanken	Informationssysteme	Sattler
3	3-8538	Datenbanken	Informationssysteme	Heuer
...

Definition

Eine Relation ist in der 3. Normalform (3NF) wenn:

- R in der 2. Normalform (2NF) ist
- kein Nichtschlüssel-Attribut transitiv von einem Kandidatenschlüssel der Relation abhängt

Definition

Jede Relation R , die nicht in 3NF ist, wird folgendermaßen zerlegt:

Die Relation $R[A, B, C]$ (auf den schnittfreien Attributmengen A, B, C) habe die irreduzierbare FD: $A \rightarrow B$, wobei A **nicht ein Schlüssel** von R ist und B ein Nichtschlüsselattribut. Dann wird durch die Zerlegung

$$R_1 = R[\underline{A}, B]$$

$$R_2 = R[\uparrow \underline{A}, C] \text{ oder } R_2 = R[\uparrow \underline{A}, \underline{C}]$$

3NF - Example



<u>ID</u>	SName	Country_Code	Area
1	Freiburg	D	357.000
2	Berlin	D	357.000
3	Orlando	USA	9.834.000
4	Bern	CH	41.285

Country Code → Area

3NF

<u>Country_Code</u>	Area
D	357.000
USA	9.834.000
CH	41.285

<u>ID</u>	SName	Country_Code (FK)
1	Freiburg	D
2	Berlin	D
3	Orlando	USA
4	Bern	CH

<u>L_ID</u>	Firma	Straße	Hausnummer	PLZ	Ort
1	Müller GmbH	Neustrasse	1	12345	Neustadt
2	Maier KG	Musterstrasse	3	34567	Musterstadt
3	Schmidt AG	Altgasse	5	98765	Altstadt
4	Mayr GbR	Schillerstrasse	8a	35781	Weilburg
5	Schneider e.K.	Pfadstrasse	5	98765	Altstadt

In welcher Normalform ist die Relation?