



# Data Warehouses

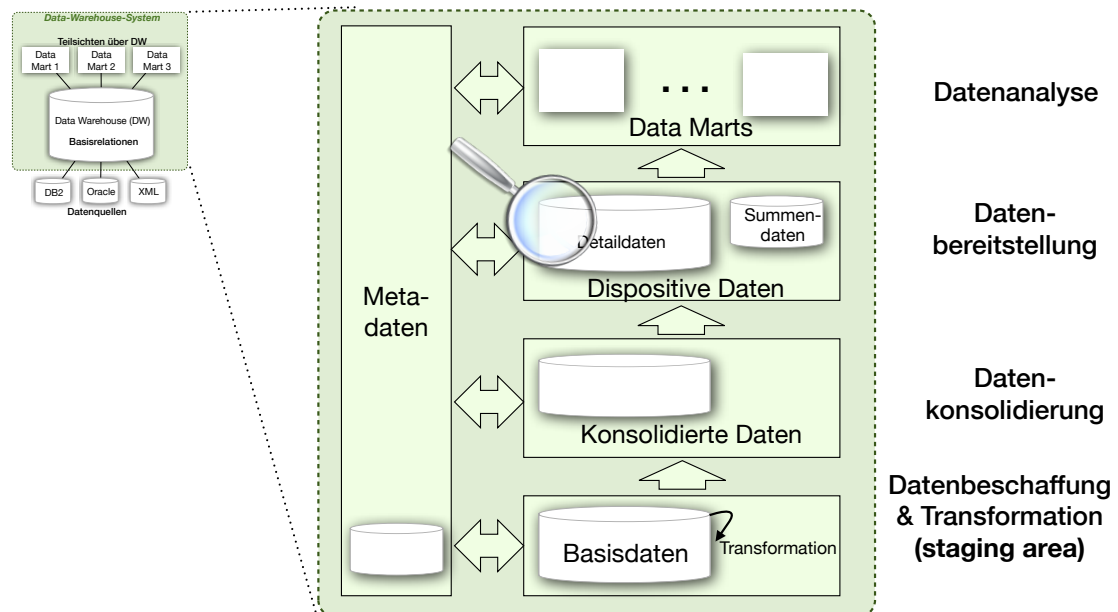
Sommersemester 2011

Melanie Herschel

[melanie.herschel@uni-tuebingen.de](mailto:melanie.herschel@uni-tuebingen.de)

Lehrstuhl für Datenbanksysteme, Universität Tübingen

## Data Warehouse Architektur

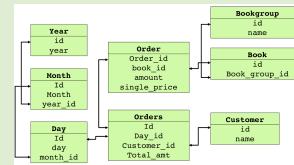
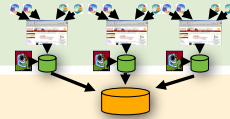


# Relationale vs. Multidimensionale Modellierung

## Schema

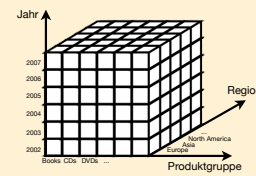
### Operative Datenbank

- Vermeidung von Redundanz / Anomalien
- Schema in 3NF
- Schema unabhängig von der Art der Anfragen entworfen



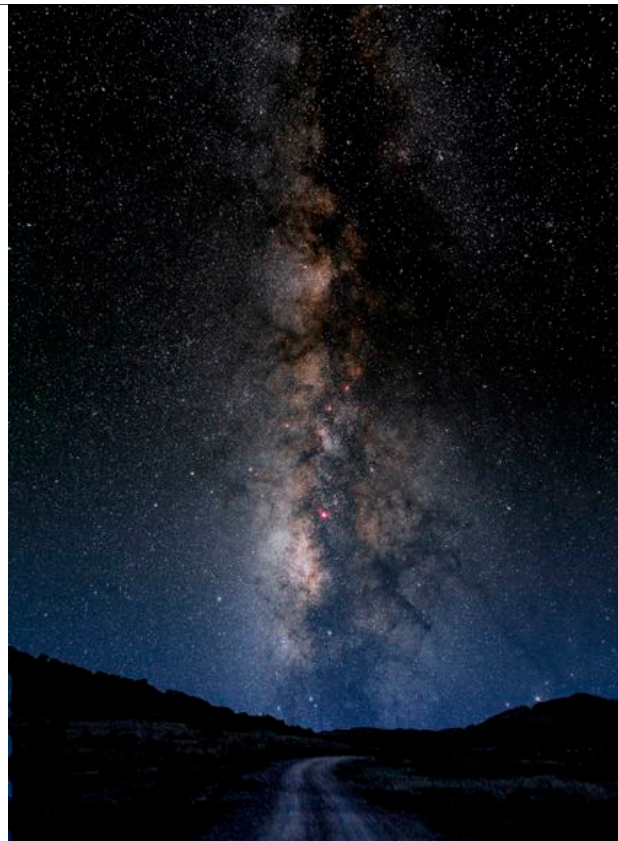
### Data Warehouse

- Modellierung von Dimensionen und Fakten, basierend auf geplante analytische Anfragen
- Redundanz möglich bzw. erwünscht
- Multidimensionales Modell (Star-Schema, Snowflake-Schema)



## Kapitel 3 Datenmodellierung

- ➔ • Konzepte & Definitionen
- Relationale Modellierung
- Modellierungsprozess



## Ausgangspunkt: Spreadsheet mit 2 Dimensionen

### Verkaufszahlen nach Zeit und Produkt

		2010												2011	
		1. Quartal			2. Quartal			3. Quartal			4. Quartal			1. Quartal	
		Jan.	Feb.	März	April	Mai	Juni	Juli	Aug.	Sept.	Okt	Nov.	Dez.	Jan.	...
Bücher	Belletristik	5	3	5	4	4	6	5	4	4	3	3	7	4 ...	
	Kinder	2	2	3	3	2	2	3	4	3	2	2	4	3 ...	
	Fachliteratur	2	2	2	2	2	2	2	3	2	2	2	1	2 ...	
Medien	Musik	5	3	5	4	4	6	5	4	4	3	3	7	4 ...	
	DVD	2	2	3	3	2	2	3	3	3	2	2	4	3 ...	
	BlueRay				2	2	2	3	4	2	2	2	4	2 ...	

Zeit-Dimension

Produkt-Dimension

## Kombination von 3 Dimensionen

### Verkaufszahlen nach Zeit und Produkt am Standort Berlin

		2010												2011	
		1. Quartal			2. Quartal			3. Quartal			4. Quartal			1. Quartal	
		Jan.	Feb.	März	April	Mai	Juni	Juli	Aug.	Sept.	Okt	Nov.	Dez.	Jan.	...
Bücher	Belletristik	5	3	5	4	4	6	5	4	4	3	3	7	4 ...	
	Kinder	2	2	3	3	2	2	3	4	3	2	2	4	3 ...	
	Fachliteratur	2	2	2	2	2	2	2	3	2	2	2	1	2 ...	
Medien	Musik	5	3	5	4	4	6	5	4	4	3	3	7	4 ...	
	DVD	2	2	3	3	2	2	3	3	3	2	2	4	3 ...	
	BlueRay				2	2	2	3	4	2	2	2	4	2 ...	

Berlin

### Verkaufszahlen nach Zeit und Produkt am Standort Stuttgart

		2010												2011	
		1. Quartal			2. Quartal			3. Quartal			4. Quartal			1. Quartal	
		Jan.	Feb.	März	April	Mai	Juni	Juli	Aug.	Sept.	Okt	Nov.	Dez.	Jan.	...
Bücher	Belletristik	5	3	5	4	4	6	5	4	4	3	3	7	4 ...	
	Kinder	2	2	3	3	2	2	3	4	3	2	2	4	3 ...	
	Fachliteratur	2	2	2	2	2	2	2	3	2	2	2	1	2 ...	
Medien	Musik	5	3	5	4	4	6	5	4	4	3	3	7	4 ...	
	DVD	2	2	3	3	2	2	3	3	3	2	2	4	3 ...	
	BlueRay				2	2	2	3	4	2	2	2	4	2 ...	

Stuttgart

## Kombination von 3 Dimensionen

		2010												2011	Berlin
		2010												2011	Stuttgart
		1. Quartal			2. Quartal			3. Quartal			4. Quartal			1. Quartal	
		Jan.	Feb.	März	April	Mai	Juni	Juli	Aug.	Sept.	Okt.	Nov.	Dez.	Jan.	...
Bücher	Belletristik	5	3	5	4	4	6	5	4	4	3	3	7	4...	
	Kinder	2	2	3	3	2	2	3	4	3	2	2	4	3...	
	Fachliteratur	2	2	2	2	2	2	2	3	2	2	2	1	2...	
Medien	Musik	5	3	5	4	4	6	5	4	4	3	3	7	4...	
	DVD	2	2	3	3	2	2	3	3	3	2	2	4	3...	
	BlueRay				2	2	2	3	4	2	2	2	4	2...	

		2010												2011	Paris
		1. Quartal			2. Quartal			3. Quartal			4. Quartal			1. Quartal	
		Jan.	Feb.	März	April	Mai	Juni	Juli	Aug.	Sept.	Okt.	Nov.	Dez.	Jan.	...
Bücher	Belletristik	5	3	5	4	4	6	5	4	4	3	3	7	4...	
	Kinder	2	2	3	3	2	2	3	4	3	2	2	4	3...	
	Fachliteratur	2	2	2	2	2	2	2	3	2	2	2	1	2...	
Medien	Musik	5	3	5	4	4	6	5	4	4	3	3	7	4...	
	DVD	2	2	3	3	2	2	3	3	3	2	2	4	3...	
	BlueRay				2	2	2	3	4	2	2	2	4	2...	

## Kombination von 3 Dimensionen

		2010												2011	Berlin
		2010												2011	Stuttgart
		1. Quartal			2. Quartal			3. Quartal			4. Quartal			1. Quartal	
		Jan.	Feb.	März	April	Mai	Juni	Juli	Aug.	Sept.	Okt.	Nov.	Dez.	Jan.	...
Bücher	Belletristik	5	3	5	4	4	6	5	4	4	3	3	7	4...	
	Kinder	2	2	3	3	2	2	3	4	3	2	2	4	3...	
	Fachliteratur	2	2	2	2	2	2	2	3	2	2	2	1	2...	
Medien	Musik	5	3	5	4	4	6	5	4	4	3	3	7	4...	
	DVD	2	2	3	3	2	2	3	3	3	2	2	4	3...	
	BlueRay				2	2	2	3	4	2	2	2	4	2...	

		2010												2011	Lyon
		1. Quartal			2. Quartal			3. Quartal			4. Quartal			1. Quartal	
		Jan.	Feb.	März	April	Mai	Juni	Juli	Aug.	Sept.	Okt.	Nov.	Dez.	Jan.	...
Bücher	Belletristik	5	3	5	4	4	6	5	4	4	3	3	7	4...	
	Kinder	2	2	3	3	2	2	3	4	3	2	2	4	3...	
	Fachliteratur	2	2	2	2	2	2	2	3	2	2	2	1	2...	
Medien	Musik	5	3	5	4	4	6	5	4	4	3	3	7	4...	
	DVD	2	2	3	3	2	2	3	3	3	2	2	4	3...	
	BlueRay				2	2	2	3	4	2	2	2	4	2...	



## Datenwürfel (*cube*)

---

- Bisher
  - Zwei Dimensionen durch ein Spreadsheet darstellbar
  - Drei Dimensionen können als Stack mehrerer 2D-Spreadsheets gesehen werden  
→ 3D Datenwürfel, engl. **cube**
- Im Allgemeinen können wir mehr als drei Dimensionen betrachten (graphisch nur schwer darstellbar).
- Auch eine Struktur mit mehr als drei Dimensionen wird *cube* (auch *hypercube*) genannt.

### (Hyper)cube

Ein **Datenwürfel**, engl. **cube** (auch **hypercube**) ist eine multidimensionale Datenstruktur, die die Speicherung und Analyse von Daten nach  $n$  Dimensionen zulässt.

Das **Schema eines  $n$ -dimensionalen Datenwürfels**  $CS$  besteht aus der Menge der **dimensionalen Schemata**  $DS$  und **Kennzahlen**  $M$  (Definitionen siehe folgende Folien), d.h.

$$CS = (DS, M) = (\{D^1, \dots, D^n\}, \{M^1, \dots, M^m\})$$

Ein Datenwürfel  $C$  ist eine Instanz eines Würfelschemas  $CS = (DS, M)$ , wobei

$$C = \text{dom}(DS) \times \text{dom}(M)$$

- Bemerkung: die Werte  $\text{dom}(DS)$  geben die Koordinaten der Werte  $\text{dom}(M)$  an.

11

## Dimensionen

---

- Zwei Anwendungen einer Dimension:
  - **Auswahl** beschreibender Daten
  - **Gruppierung** beschreibender Daten im gewünschten Detailgrad
- Eine Dimension wird als **containment-hierarchy** definiert.
- Diese Hierarchie hat mehrere **Ebenen** (*levels*) die jeweils einen für Analysen relevanten Detailgrad beschreiben.
- Die **oberste Ebene** (Wurzel) beschreibt die gesamte Dimension.
- Manche Hierarchie speichert auch **level properties**, die einfache, nicht-hierarchische Informationen pro Ebene speichern (z.B. Anzahl Einwohner pro Stadt in Ortdimension)

12

# Dimensionen

## Schema einer Dimension

Das Schema einer Dimension  $D$  besteht aus einer partiell geordneten Menge von Kategorieattributen  $\{\{D_1, \dots, D_n, Top_D; \rightarrow\}\}$ , wobei

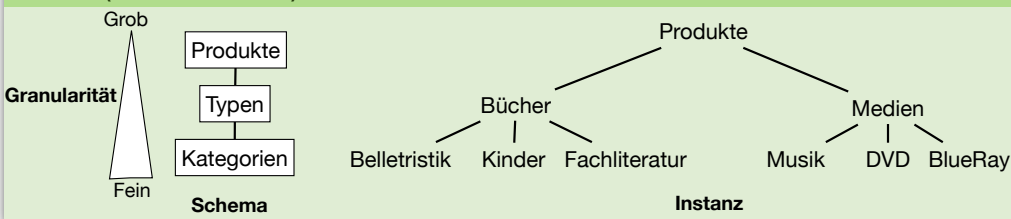
- $\rightarrow$  die funktionale Abhängigkeit bezeichnet und
- $Top_D$  ein generisches maximales Element in Bezug auf  $\rightarrow$  darstellt, so dass  $Top_D$  von allen Attributen funktional bestimmt wird, d.h.

$$\forall i (1 \leq i \leq n), D_i \rightarrow Top_D.$$

Des Weiteren existiert genau ein  $D_i$ , welches alle anderen Kategorieattribute bestimmt und somit die feinste Granularität einer Dimension vorgibt, d.h.

$$\exists i (1 \leq i \leq n) \forall j (1 \leq j \leq n, i \neq j): D_i \rightarrow D_j$$

## Hierarchie (Schema und Instanz) der Produktdimension

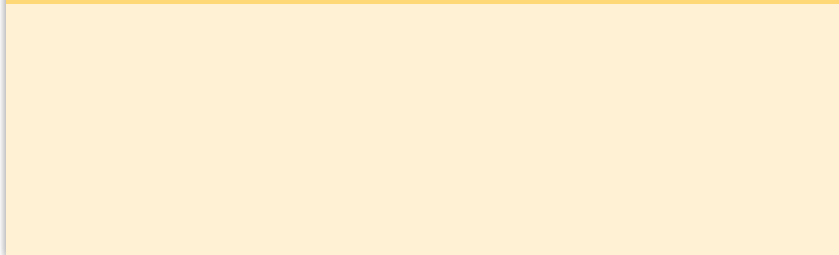


13

# Dimensionen

		2010												2011	
		1. Quartal			2. Quartal			3. Quartal			4. Quartal			1. Quartal	
		Jan.	Feb.	März	April	Mai	Juni	Juli	Aug.	Sept.	Okt.	Nov.	Dez.	Jan.	...
Bücher	Belletristik	5	3	5	4	4	6	5	4	4	3	3	7	4	...
	Kinder	2	2	3	3	2	2	3	4	3	2	2	4	3	...
	Fachliteratur	2	2	2	2	2	2	2	3	2	2	2	1	2	...
Medien	Musik	5	3	5	4	4	6	5	4	4	3	3	7	4	...
	DVD	2	2	3	3	2	2	3	3	3	2	2	4	3	...
	BlueRay				2	2	2	3	4	2	2	2	4	2	...

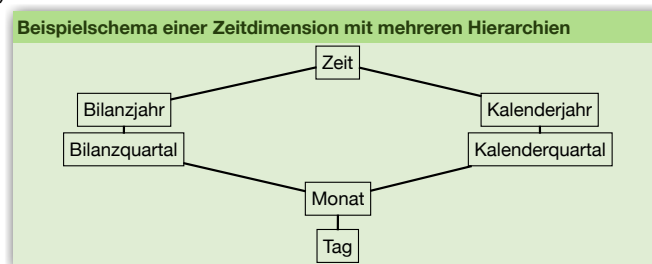
## Hierarchie (Schema und Instanz) der Zeitdimension



14

# Dimensionen

- Grundsätzlich gibt es **keine spezifische Reihenfolge** der dimensionalen Werte.
  - Aber möglich, z.B. Zeit-Dimension, wo Werte laut Zeitachse sortiert werden.
- Einzige notwendige Strukturierung ist die **Containment-Beziehung** von Werten der Ebene  $i$  in Werten der Ebene  $i+1$ 
  - Z.B. Musik, DVD, BlueRay auf Ebene 1  $\in$  Medien auf Ebene 2
- Prinzipiell sind auch **mehrere Hierarchien pro Dimension möglich**.
  - z.B. Kalenderjahr und Bilanzjahr
  - Diese teilen sich ein oder mehrere unterste Ebenen (Ebene 0, Ebene 0+1, Ebene 0+1+2, ...) und definieren unterschiedliche höhere Ebenen.



Data Warehouses | SS 2011 | Melanie Herschel | Universität Tübingen

15

# Dimensionen

- Häufige Annahmen:
  - Verwendung **balancierter Hierarchien**
    - Jeder Pfad von der Wurzel zu einem Blattknoten hat die gleiche Länge.
  - Auf der Instanzebene einer Hierarchie können keine Ebenen übersprungen werden, es sind **nur direkte Eltern-Kind-Verknüpfungen** möglich.
    - Gibt es z.B. Städte, die Bundesländern zugeordnet sind, so muss jede Stadt einem Bundesland zugeordnet werden (auch Stadtstaaten wie Berlin, oder Städte wie Washington DC, die keinem Bundesland zugehören).
  - Es gibt **genau einen Elternwert pro Kindwert** in einer Hierarchie (siehe Definition von Bäumen).
    - Ein Produktwert kann nicht zwei Produktkategorien untergeordnet werden.
- In der Praxis gibt es Möglichkeiten, diese Annahmen zu lockern, diese besprechen wir in dieser Vorlesung nicht. Wir nehmen stets an, dass diese Annahmen erfüllt sind.

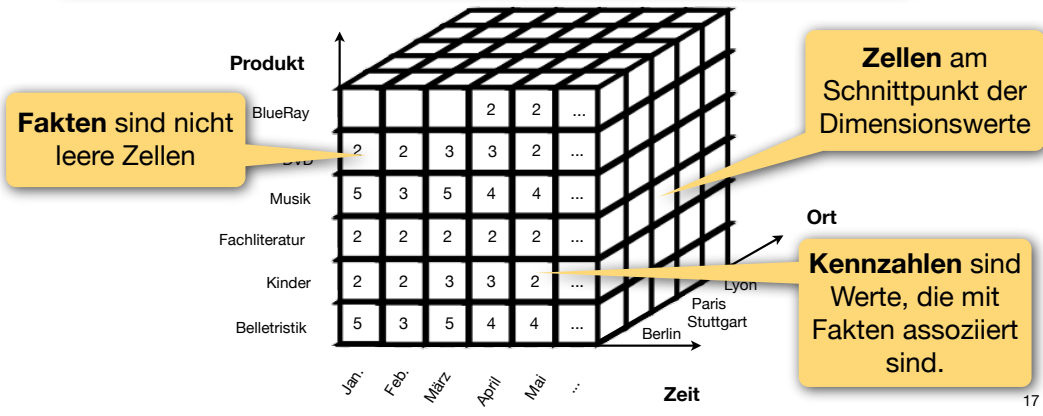
Data Warehouses | SS 2011 | Melanie Herschel | Universität Tübingen

16

# Zellen (cells), Fakten (facts), Kennzahlen (measures) Überblick

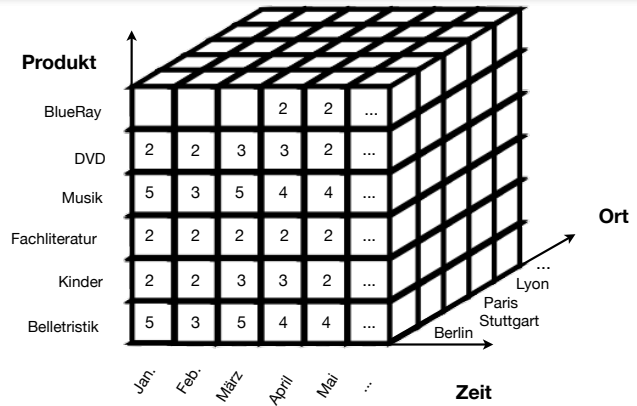
		2010												2011
		1. Quartal			2. Quartal			3. Quartal			4. Quartal			1. Quartal
		Jan.	Feb.	März	April	Mai	Juni	Juli	Aug.	Sept.	Okt.	Nov.	Dez.	Jan.
Bücher	Belletristik	5	3	5	4	4	6	5	4	4	3	3	7	4 ...
	Kinder	2	2	3	3	2	2	3	4	3	2	2	4	3 ...
	Fachliteratur	2	2	2	2	2	2	2	3	2	2	2	1	2 ...
Medien	Musik	5	3	5	4	4	6	5	4	4	3	3	7	4 ...
	DVD	2	2	3	3	2	2	3	3	3	2	2	4	3 ...
	BlueRay				2	2	2	3	4	2	2	2	4	2 ...

Berlin



# Zellen (cells), Fakten (facts), Kennzahlen (measures) Überblick

Interpretation von Fakten & Interpretation leerer Zellen



# Zellen (*cells*), Fakten (*facts*), Kennzahlen (*measures*)

## Fakten

- Fakten sind die Objekte, die die Subjekte der geplanten Analysen beschreiben.
  - Z.B., Verkaufszahlen, Umsätze, ...
- Fakten werden implizit durch ihre Dimensions-Kombination definiert
  - Z.B. Verkaufszahlen nach Monat, Kategorie und Standort
- Existiert eine nicht leere Zelle für eine Dimensions-Kombination, so existiert ein Fakt für diese Kombination; sonst nicht.
- Fakten haben eine **Granularität**, die den Detailgrad der Information beschreibt.
- Die Granularität wird durch die **Assoziation eines Fakts mit einer Ebene der Dimensionshierarchien** bestimmt.
  - Die Granularität ist **feiner**, je näher diese Ebene der Blattebene.  
Z.B. Kategorie pro Monat pro Stadt
  - Die Granularität ist **gröber**, je näher diese Ebene der Wurzel.  
Z.B. Produkt pro Jahr pro Land

19

# Zellen (*cells*), Fakten (*facts*), Kennzahlen (*measures*)

## Fakten

### Ereignis-Fakt (*event fact*)

- Modelliert **Ereignisse der realen Welt** (zumindest auf der Ebene der feinsten Granularität).
- Es existiert genau ein Fakt für jedes bestimmte Ereignis des reale-Welt-Prozesses.
- Ereignisse können im Prinzip **unabhängig** und **zu jeder Zeit** auftreten.

### Beispiele von Ereignis-Fakten

- Ein Fakt für jeden Verkauf eines bestimmten Buchs (feine Granularität)
- Ein Fakt für jeden Tag, an dem mindestens eine Kopie eines bestimmten Buchs verkauft wurde (grobe Granularität)
- Ein Verkauf (Fakt) ist genau an einen Zeitpunkt, Ort und ein Produkt gebunden.

### Snapshot-Fakt (*snapshot fact*)

- Modelliert den **aktuellen Status** eines Prozesses.
- Das gleiche Objekt (mit dem sich der Prozess befasst) kann in mehreren Fakten zu verschiedenen Zeitpunkten auftreten.
- Wird oft **periodisch** erfasst.

### Beispiele von Ereignis-Fakten

- Lagerbestand pro Produkt pro Lager.
- Das gleiche Produkt kann zu mehreren Fakten beitragen, da z.B. die gleiche CD sowohl im Mai als auch im Juni auf Lager sein kann und somit zu beiden Lagerbeständen beiträgt.

20

## Zellen (*cells*), Fakten (*facts*), Kennzahlen (*measures*)

### Kennzahlen

---

- Eine **Kennzahl** beschreibt einen Fakt und kann auf Kennzahlen anderer Fakten basieren.
- Daher hat eine Kennzahl zwei Bestandteile
  - Eine **numerische Eigenschaft** des beschriebenen Fakts  
z.B. Verkaufspreis, Profit, ...
  - Eine **Formel** (auch Berechnungsvorschrift) **zur Kombination mehrerer Kennzahlen**
    - Skalarfunktionen, z.B. Umsatzsteueranteil = Menge x Preis x Steuersatz
    - Aggregationsfunktionen, z.B. SUM, AVG, Standardabweichung
    - Ordnungsbasierte Funktionen, z.B. Kumulation, Top-k Berechnung
- Eine Kennzahl hat stets einen **numerischen Datentyp**.
- Im Allgemeinen kann **mehr als eine Kennzahl pro Zelle** gesammelt werden.
  - z.B. Anzahl Verkäufe und Gesamtumsatz pro Monat, Kategorie und Standort

## Zellen (*cells*), Fakten (*facts*), Kennzahlen (*measures*)

### Kennzahlen

---

- Genau wie Fakten besitzen auch Kennzahlen eine gewisse **Granularität**.
- Frage: wie leitet man eine Kennzahl für eine andere Granularität (= Kennzahl eines Fakts anderer Granularität) ab?
- Antwort: auch dafür wird eine der Kennzahl zugeordnete Formel verwendet.
- Aber: Dies ist nicht immer möglich, wir unterscheiden drei Möglichkeiten:
  - **Additive Kennzahlen:** Kennzahlen, die entlang jeder Dimension aggregiert werden können.  
Z.B. macht es Sinn, Verkaufszahlen über alle drei Dimensionen Ort, Zeit und Produkt zu summieren (daraus ergibt sich der Gesamtverkauf).
  - **Semi-additive Kennzahlen:** Kennzahlen, die mindestens entlang einer Dimension nicht aggregiert werden können.  
Oft der Fall bei Snapshot-Fakten, z.B. macht es keinen Sinn, Lagerbestand pro Monat aufzusummieren um Jahresbestand zu berechnen (das Ergebnis entspricht keinem Fakt).
  - **Nicht-additive Kennzahlen:** Kennzahlen, deren Aggregation entlang keiner Dimension Sinn macht.  
Z.B. lässt sich ein Durchschnittswert auf Ebene  $i$  nicht durch Durchschnittswerte auf Ebene  $i+1$  berechnen, egal entlang welcher Dimension.

## Kapitel 3

# Datenmodellierung

---

- Konzepte & Definitionen
- ➔ • Relationale Modellierung
- Modellierungsprozess

23



## Annahmen

---

- Ziel ist es, einen Datenwürfel, assoziierte Fakten und Kennzahlen im **relationalen Datenmodell** darzustellen.
- Gegeben:
  - Schema eines Datenwürfels  $CS = (DS, M)$ , wobei
    - $DS = \{D^1, \dots, D^n\}$  die Menge der  $n$  Dimensionsschemata
    - $M = \{N^1, \dots, M^m\}$  die Menge der Kennzahlen

24

# Star Schema

## Definition

### Star Schema

- Ein **Star Schema** wird durch eine **Menge von Dimensionstabellen** und einer **Faktentabelle** definiert.
- **Dimensionstabellen:** Für jede Dimension  $D^i \subseteq DS$  mit Schema  $(D_1, \dots, D_k, TopD)$  existiert eine Tabelle mit dem relationalen Schema

$$D^i(PK, D_1, \dots, D_k)$$

wobei  $PK$  ein Primärschlüssel ist und jedes  $D_j$  einer Ebene des hierarchischen Schemas  $D^i$  (ausgenommen der obersten Ebene  $TopD$ ) entspricht.

- **Faktentabelle:** die Faktentabelle  $F$  entspricht dem Schema

$$F(FK_1 \rightarrow D_1.PK, \dots, FK_n \rightarrow D_n.PK, M^1, \dots, M^m)$$

das einen Fremdschlüssel  $FK_i$  zu jeder der  $n$  Dimensionstabellen beinhaltet und ein Attribut pro Kennzahl speichert. Die Kombination aller Fremdschlüssel ergibt den Primärschlüssel der Faktentabelle.

25

# Star Schema

## Beispiel

### Beispiel eines Star Schemas

#### Produkt\_Dimension

PID	Kategorie	Typ
1	Belletristik	Bücher
2	Kinder	Bücher
3	Fachliteratur	Bücher
4	Musik	Medien
5	DVD	Medien
6	BlueRay	Medien

#### Zeit\_Dimension

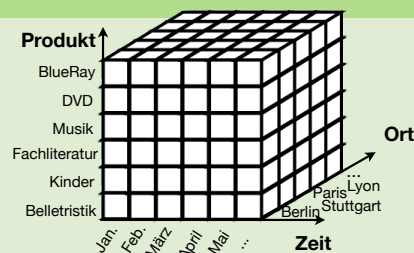
ZID	Monat	Quartal	Jahr
1	Jan10	Q1 2010	2010
2	Feb10	Q1 2010	2010
...	...	...	...

#### Ort\_Dimension

OID	Stadt
1	Berlin
2	Stuttgart
3	Paris
4	Lyon

#### Faktentabelle

PID	ZID	OID	#Verkäufe	Gewinn
1	1	1	5	30
1	1	2	5	37
1	1	3	5	45
1	1	4	5	20
2	1	1	2	33
2	1	2	2	35
2	1	3	2	40
2	1	4	2	35
...	...	...	...	...
1	2	1	3	22
...	...	...	...	...



Kennzahlen (Verkäufe & Gewinn) für Belletristik im Januar 2010 in Stuttgart.

Alle weiteren Kombinationen von Produktkategorien und Orten im Januar 2010.

Beginn der Kombinationen für Februar 2010 (danach auch für alle weiteren Monate)

6

# Star Schema

## Bemerkungen

- Redundanz in Dimensionstabellen, die typischerweise in 2NF gehalten werden.
- Im Vergleich zur Faktentabelle sind die Dimensionstabellen klein, daher führt die Redundanz nicht zu signifikanten Speicherverschwendungen.
- Die Faktentabelle weist die 3NF auf.
- Es ist kein Attribut für die Wurzeldimension  $Top_D$  nötig, da die Werte in allen Tupeln der entsprechenden Dimensionstabelle gleich wären.
- Werte des Schlüssel einer Dimensionstabelle sind üblicherweise generierte Werte (*surrogate keys*) ohne bestimmte Semantik.
  - Weniger Speicherbedarf, als z.B. eine ISBN als ProduktID
  - Schnellere Anfragebearbeitung

27

# Snowflake Schema

## Definition

### Snowflake Schema

- Ein **Snowflake Schema** wird durch eine **Menge von Dimensionstabellen** und einer **Faktentabelle** definiert.

- **Dimensionstabellen:** Für jede Dimension  $D^i \subseteq DS$  mit Schema  $(D_1, \dots, D_k, Top_D)$  existieren  $k$  Tabellen mit dem relationalen Schema

$$D_j^i(PK, A_1, \dots, A_m, FK_{j+1} \rightarrow D_{j+1}^i) \text{ für } 1 \leq j < k$$
$$D_k^i = (PK, A_1, \dots, A_m)$$

wobei  $PK$  ein Primärschlüssel ist und jedes  $D_j^i$  einer Ebene des hierarchischen Schemas  $D^i$  entspricht.  $FK_j$  ist ein Fremdschlüssel auf die Tabelle  $D_j^i$ , wobei  $1 < j \leq k$ .  $A_1, \dots, A_m$  sind textuelle Attribute zur Beschreibung relevanter Daten der aktuellen Ebene.

- **Faktentabelle:** die Faktentabelle  $F$  entspricht dem Schema

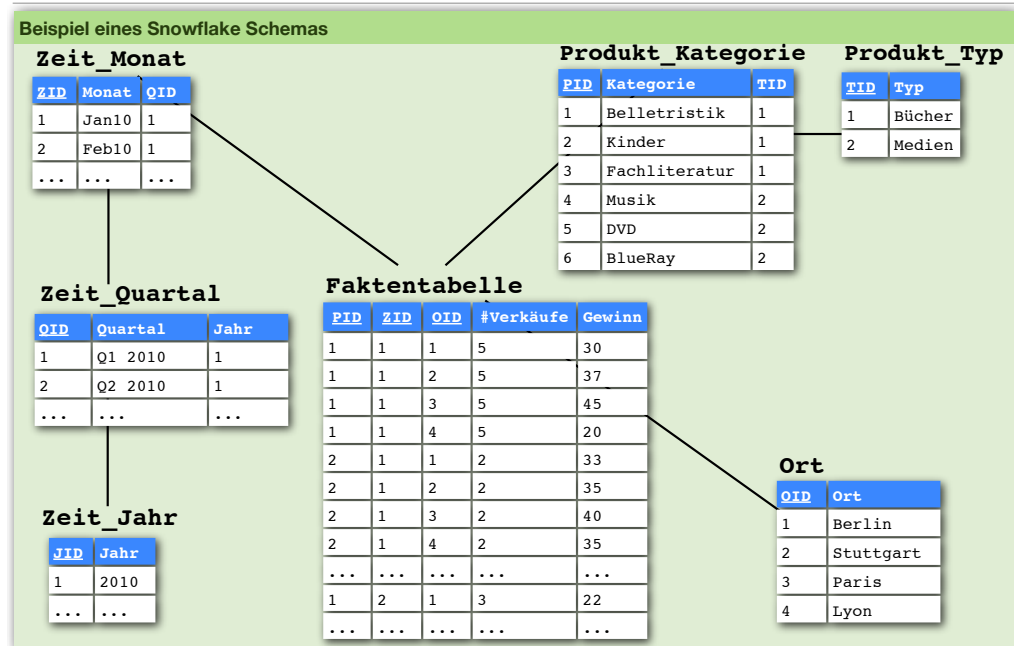
$$F(FK_1 \rightarrow D_1^1.PK, \dots, FK_n \rightarrow D_n^1.PK, M^1, \dots, M^m)$$

das einen Fremdschlüssel  $FK_i$  zu jeder der  $n$  Dimensionstabellen feinsten Granularität beinhaltet und ein Attribut pro Kennzahl speichert. Die Kombination aller Fremdschlüssel ergibt den Primärschlüssel der Faktentabelle.

28

# Snowflake Schema

## Beispiel



Data Warehouses | SS 2011 | Melanie Herschel | Universität Tübingen

29

# Snowflake Schema

## Bemerkungen

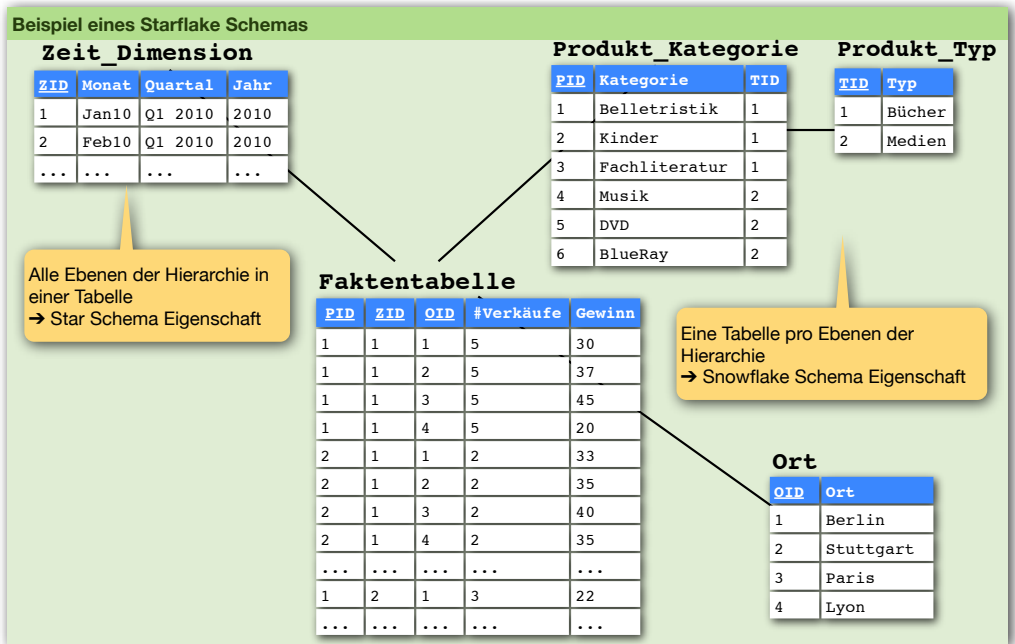
- Redundanz in Dimensionstabellen durch Normalisierung vermieden.
- Anfragebearbeitung schwieriger, da mehr Joins notwendig werden.
- Mehrere Dimensionstabellen liefern eine explizite Darstellung der Dimensionshierarchien.
- Wahl zwischen Star und Snowflake Schema hängt stark von den Anforderungen an die Anwendung ab.
- Auch eine Mischform von Star und Snowflake Schema (Starflake Schema) ist möglich, d.h., einige Dimensionen in 3NF, andere in 2NF.

Data Warehouses | SS 2011 | Melanie Herschel | Universität Tübingen

30

# Starflake Schema

## Beispiel



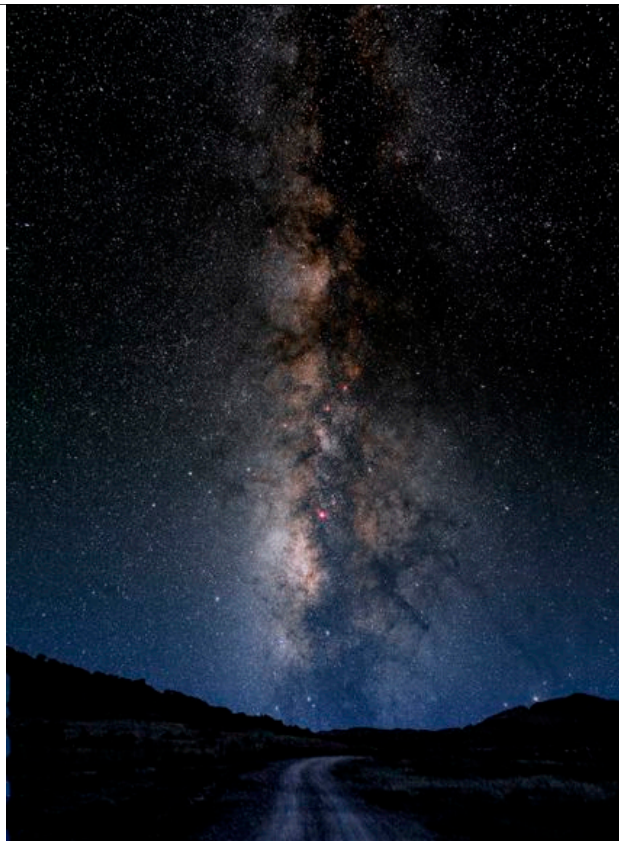
Alle Ebenen der Hierarchie in einer Tabelle  
→ Star Schema Eigenschaft

Eine Tabelle pro Ebenen der Hierarchie  
→ Snowflake Schema Eigenschaft

## Kapitel 3

### Datenmodellierung

- Konzepte & Definitionen
- Relationale Modellierung
- ➔ • Modellierungsprozess



# Allgemeine Designprinzipien

---

- Zwei wesentliche Unterschiede zur Datenmodellierung im “klassischen” relationalen Modell.
- Das Modell sollte nicht versuchen, alle möglichen / existierenden Daten und Beziehungen darzustellen. Nur die **für Analysen wichtige Informationen** sollten modelliert werden.
- **Redundanz** ist an wenigen, ausgewählten Stellen (Dimensionstabellen) **akzeptabel**.

# Designprozess

---

Designprozess (nach Kimball) in vier Schritten:

1. Wähle Geschäftsprozess(e) aus, die zu modellieren sind.
2. Wähle die Granularität des Geschäftsprozesses.
3. Entwerfe die Dimensionen.
4. Wähle die Kennzahlen.

# Designprozess

## Beispiel

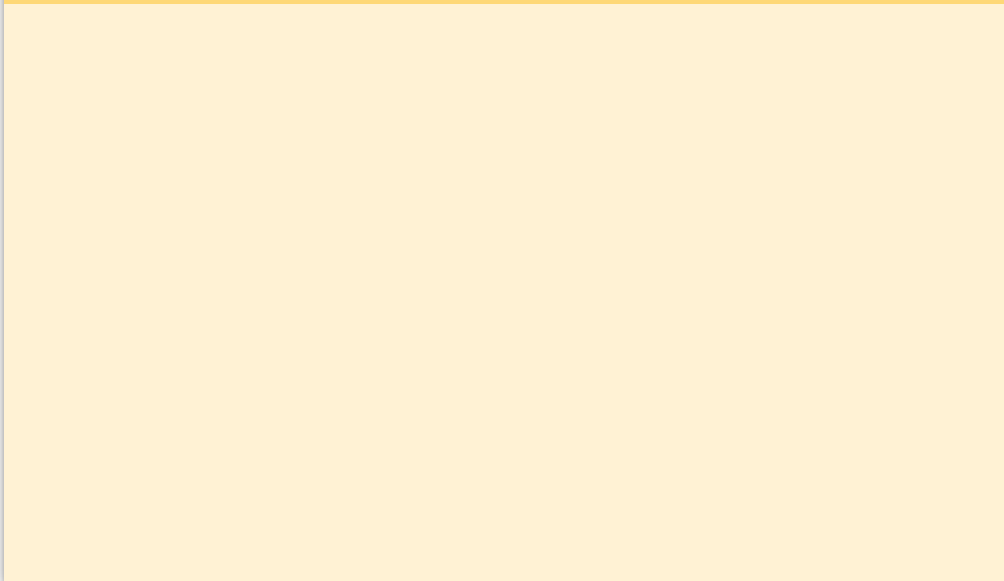
### Designprozess im Fall eines Buchhandels

- Schritt 1
  - Es existiert ein Geschäftsprozess für (i) Buchverkäufe an Kunden und (ii) Buchkäufe von Verlagen.
  - Prozess (i) ist vermutlich der relevanteste wenn es darum geht, den Profit zu erhöhen.  
➔ Wahl des Buchverkaufsprozess (ii)
- Schritt 2
  - (i) Granularität Einzelverkauf pro Buch pro Filiale vs. (ii) Granularität Gesamtverkauf eines Buchs pro Filiale pro Tag.
  - Granularität (ii) ausreichend um Filialen und Bücher zu bewerten, spart Speicherplatz und beschleunigt die Bearbeitung relevanter Anfragen.  
➔ Wahl der gröberen Granularität (ii)
- Schritt 3: Spezifikation der Dimensionen für Filialen, Bücher und Datum.
- Schritt 4: Definition relevanter Kennzahlen, z.B. Anzahl Verkäufe, Umsatz, Kosten, Gewinn.

# Designprozess

## Beispiel

### Definition der Dimensionshierarchien und Entwicklung eines entsprechenden Snowflake Schemas



# Zusammenfassung

---

- Wichtige Konzepte
  - Datenwürfel
  - Dimensionen
  - Fakten
  - Kennzahlen
- Relationale Modellierung
  - Star Schema
  - Snowflake Schema
- Entwicklungsprozess in vier Schritten

