Data-Warehouse-Technologien

Prof. Dr.-Ing. Kai-Uwe Sattler[1]

Prof. Dr. Gunter Saake[2]
Dr. Veit Köppen[2]

[1]TU Ilmenau
FG Datenbanken & Informationssysteme

[2]Universität Magdeburg
Institut für Technische und Betriebliche Informationssysteme

Letzte Änderung: 18.10.2019

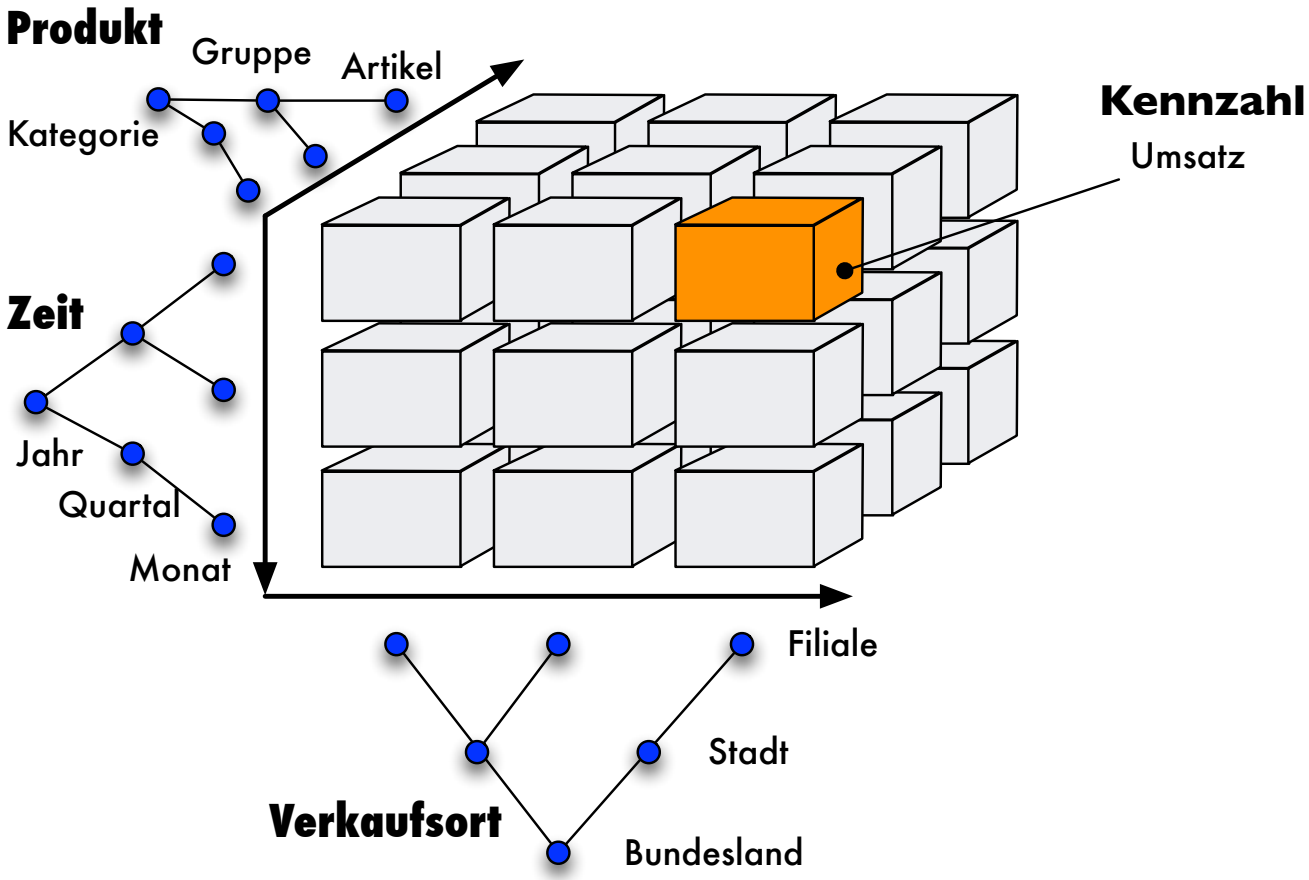
Teil III

Multidimensionales Datenmodell

Multidimensionales Datenmodell

Grundbegriffe

- Dimensionen
- Fakten / Kennzahlen



Motivation

- Datenmodell ausgerichtet auf Unterstützung der Analyse
- Datenanalyse im Entscheidungsprozess
  - Betriebswirtschaftliche Kennzahlen stehen im Mittelpunkt
    $\rightarrow$ Fakten
    • Gewinn, • Umsatz, • Kosten, • etc.
  - Betrachtung der Kennzahlen aus unterschiedlichen Perspektiven
    $\rightarrow$ Dimensionen
    • Zeit, • Raum, • Sache
  - Unterteilung der Auswertedimensionen möglich
    $\rightarrow$ Hierarchien oder Konsolidierungsebenen
    • Jahr, • Quartal, • Monat

Verfügbare Informationen

- Qualifizierend
  - Repräsentiert durch „Kategorienattribute“
  - Daten zur Nutzung als Navigationsraster („Drill-Pfade“)
  - Modelliert als Begriffshierarchien im Rahmen von Dimensionen
- Quantifizierend
  - Bilden Gegenstand der Auswertung („Summenattribute“ oder andere arithmetische Operationen)
  - Zellen eines Würfels, mit Dimensionen als Kanten

Dimensionen

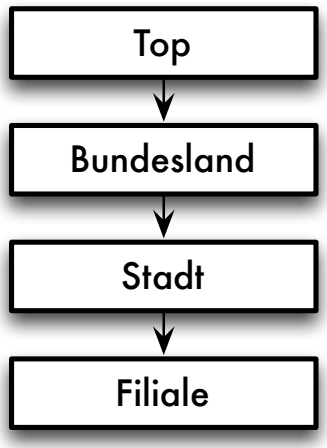
- Dimension:
  - Beschreibt mögliche Sicht auf assoziierte Kennzahlen
  - Endliche Menge von n ($n \geq 2$) Dimensionselementen (Hierarchieobjekten), die eine semantische Beziehung aufweisen
  - Dienen der orthogonalen Strukturierung des Datenraums
- Beispiele:
  - Produkt,
  - Filialstruktur,
  - Geschäftsjahr

Hierarchien in Dimensionen

- Dimensionselemente:
  - Knoten einer Klassifikationshierarchie
  - Klassifikationsstufe beschreibt Verdichtungsgrad
  - Darstellung von Dimensionen über Klassifikationsschema (Schema von Klassifikationshierarchien)
- Formen:
  - Einfache Hierarchien
  - Parallele Hierarchien

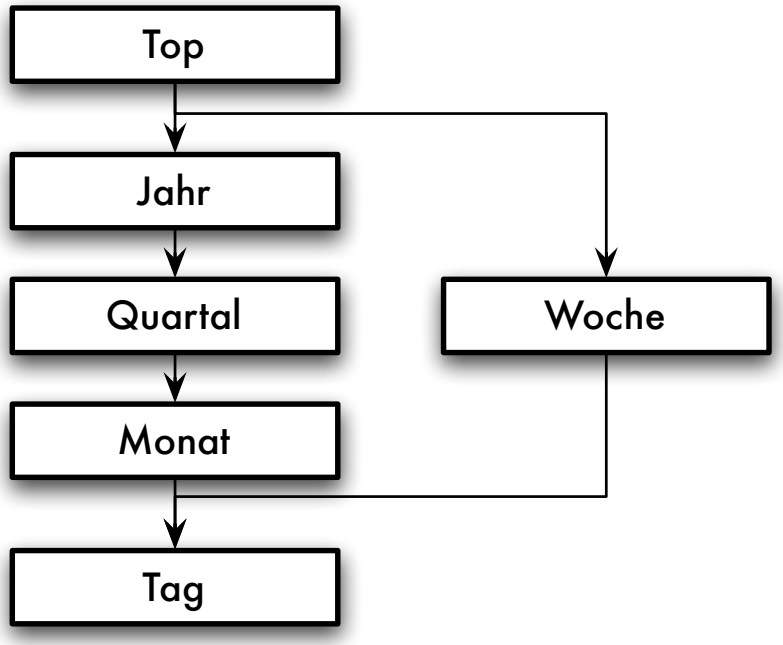
Einfache Hierarchie

- Höhere Hierarchieebene enthält die aggregierten Werte genau einer niedrigeren Hierarchiestufe
- Oberster Knoten: *Top*
  - Enthält Verdichtung auf einen einzelnen Wert für die Dimension



Parallele Hierarchie

- Innerhalb einer Dimension sind mehrere unabhängige Arten der Gruppierung möglich
- Keine hierarchische Beziehung zwischen parallelen Zweigen
- Parallelhierarchie
  - Pfad im Klassifikationsschema
  - Konsolidierungspfad



Schema einer Dimension D

- Partiell geordnete Menge von Kategorienattributen $(\{D^1, \ldots, D^n, Top_D\}; \rightarrow)$
  - Generisches maximales Element Top_D
  - Funktionale Abhängigkeit $\rightarrow$
- Top_D wird von allen Attributen funktional bestimmt:

$$\forall i, 1 \leq i \leq n : D_i \rightarrow Top_D$$

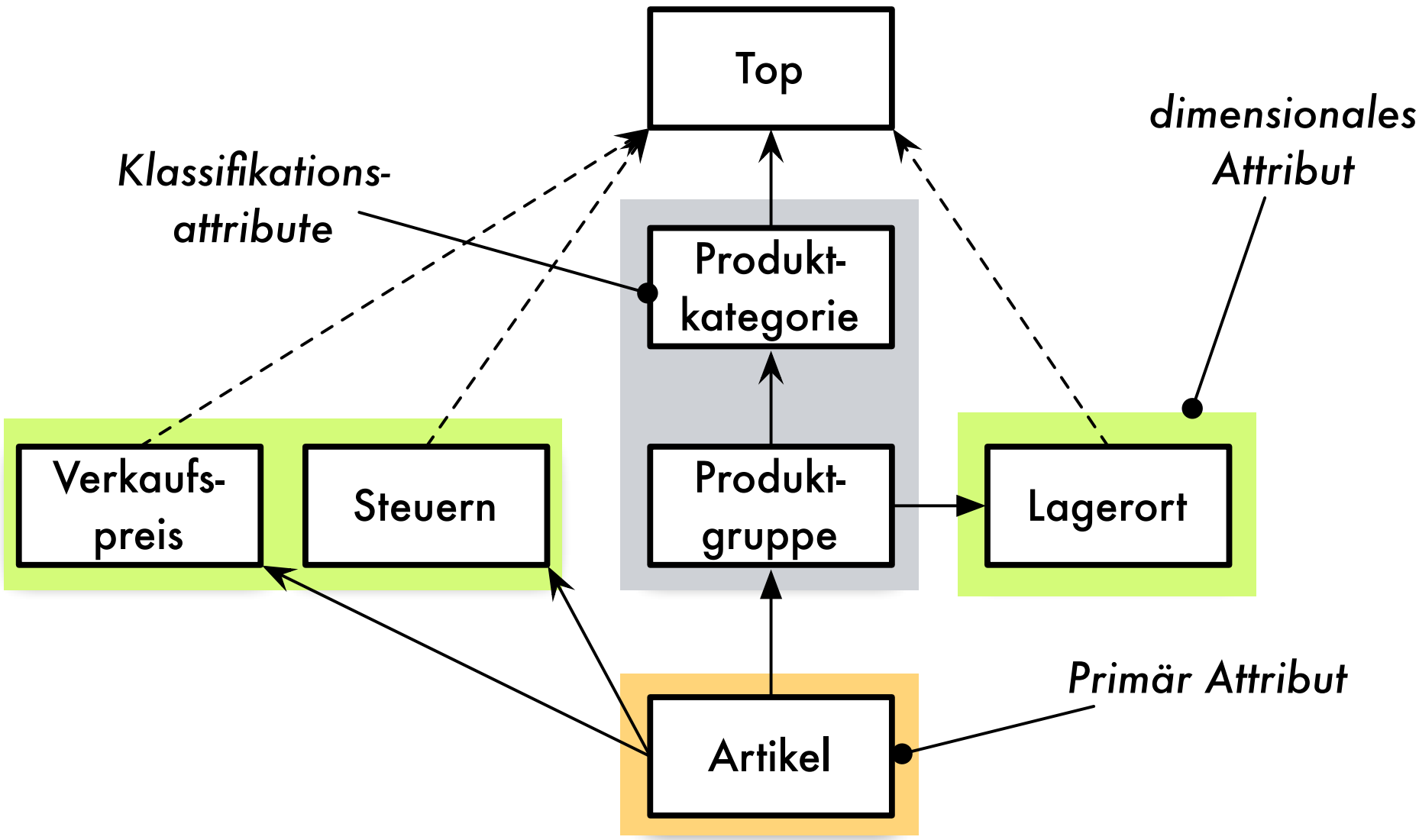
- Es gibt genau ein D_i, das alle anderen Kategorieattribute bestimmt
  - Gibt feinste Granularität einer Dimension vor

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

Kategorienattribute

- Inhaltliche Verfeinerung durch unterschiedliche Rollen:
- Primärattribut
  - Kategorienattribut, das alle anderen Attribute einer Dimension bestimmt
  - Definiert maximale Feinheit
  - Beispiel: „Auftragsposition“
- Klassifikationsattribut
  - Element der Menge, die mehrstufige Kategorisierung (Klassifikationshierarchie) bilden
  - Beispiel: „Produkt“, „Produktgruppe“, „Produktkategorie“
- Dimensionales Attribut
  - Element der Menge der Attribute, die vom Primärattribut oder einem Klassifikationsattribut bestimmt werden und nur Top_D bestimmen
  - Beispiel: „Regalposition“

Struktur einer Dimension: Beispiel



Kennzahlen

- Kennzahlen / Fakten (engl. facts):
  - (Verdichtete) numerische Messgrößen
  - Beschreiben betriebswirtschaftliche Sachverhalte
- Fakt: Maßzahl (engl. measure)
- Kennzahl:
  - Aus Fakten konstruiert (abgeleitete Kennzahl)
  - Durch Anwendung arithmetischer Operationen
- Beispiele:
  - Umsatz, Gewinn, Kosten
  - Deckungsbeitrag, ROI (Return on Investment)
  - Fluktuationsquote, Umsatzsteigerung

Fakt: Schema

- Schema wird durch mehrere Bestandteile spezifiziert
- Granularität $G = \{G_1, \ldots, G_k\}$
  - G ist Teilmenge aller Kategorienattribute aller im Schema existierenden Dimensionsschemata $DS_1, \ldots, DS_n$
    - ★ $\forall i, 1 \leq i \leq k, \exists j, 1 \leq j \leq n : G_i \in DS_j$
    - ★ $\forall i, 1 \leq i \leq k, \forall j, 1 \leq j \leq k, i \neq j : G_i \not\rightarrow G_j$
      (keine funktionalen Abhängigkeit zwischen Kategorienattributen einer Granularität)
  - „Detailliertheitsgrad“ der Fakten
- Summationstyp $SumTyp$
  - Bestandsgröße
  - Stromgröße
  - Einheit

Kennzahl

- Kennzahl M ist definiert durch
  - Granularität G
  - Berechnungsvorschrift $f()$ über Fakten
  - Summationstyp $SumTyp$
- Berechnung über nichtleerer Teilmenge der im Schema existierenden Fakten
- $M = (G, f(F_1, \ldots, F_k), SumTyp)$

Kennzahl: Bildung von f()

- Skalarfunktionen
  - $+, -, *, /, mod$
  - Beispiel: $Umsatzsteueranteil = Menge * Preis * Steuersatz$
- Aggregatfunktionen
  - Funktion $H()$ zur Verdichtung eines Datenbestandes, indem aus n Einzelwerten ein Aggregatwert ermittelt wird

$$H : 2^{dom(X_1) \times \cdots \times dom(X_n)} \rightarrow dom(Y)$$

  - Bsp.: $SUM(), AVG(), MIN(), MAX(), COUNT()$
- Ordnungsbasierte Funktionen
  - Definition von Kennzahlen auf Basis zuvor definierter Ordnungen
  - Bsp.: Kumulation, $TOP(n)$, $MEDIAN()$

Summationstypen

- Zuweisung eines Summationstyps charakterisiert erlaubte Aggregationsoperationen
- $FLOW$
  - zeitraumbezogen (pro Zeiteinheit)
  - Beliebig aggregierbar
  - Beispiel: Bestellmenge eines Artikels pro Tag
- $STOCK$
  - Maß über Zeitraum
  - Beliebig aggregierbar mit Ausnahme temporaler Dimension
  - Beispiel: Lagerbestand, Einwohnerzahl
- $VALUE-PER-UNIT(VPU)$
  - zeitpunktbezogen (zum Zeitpunkt)
  - Aktuelle Zustände, die nicht summierbar sind
  - Zulässig nur: $MIN(), MAX(), AVG()$
  - Beispiele: Preis, Wechselkurs, Steuersatz

Summierbarkeit

	FLOW	*STOCK*		*VPU*
		Aggregation über temporale Dimension?		
		nein	ja	
MIN/MAX	+	+		+
SUM	+	+	−	−
AVG	+	+		+
COUNT	+	+		+

Disjunktheit

- Konkreter Wert einer Kennzahl geht exakt einmal in Ergebnis ein
- Bsp.: Verkäufe einer Filiale

Umsatz	**2010**	**2011**
Bier	38	42
Biermix	27	31
Softdrinks	54	57
Gesamt	92	99

- Wie ist der Gesamtumsatz pro Jahr?
- Wie ist der Gesamtumsatz über die Produktgruppen im Jahr?

Vollständigkeit

- Kennzahlen auf höherer Aggregationsebene lassen sich komplett aus Werten tieferer Stufen berechnen

Weinregion	**2010**	**2011**
Rheinhessen	152	153
Saale-Unstrut	98	104
Mosel	161	172
Sonstige	20	22
Gesamt	431	451

- Sonstige = Einige Produzenten aus anderen Gebieten

Simpson Paradox

- Gruppenbewertungen ergeben unterschiedliche Ergebnisse hinsichtlich der betrachteten Aggregationsstufe
- Quotientenbildung auf stark unterschiedlichen Gruppengrößen

	Rotwein			Weißwein		
	mit Prädikat	Gesamt Gesamt	Prädikats-quote	mit Prädikat	Gesamt Gesamt	Prädikats-quote
2010	5	5	100%	7	8	88%
2011	10	15	67%	1	2	50%
Summe	15	20	75%	8	10	80%

Aggregatfunktionen

X ist Klassifikationsknoten und $(X_1, X_2, \ldots, X_n)$ ist Partitionierung

- Distributive Aggregatfunktion: $\exists g : f(X) = f\left(g(X_1), g(X_2), \ldots, g(X_n)\right)$
- Algebraische Aggregatsfunktion: f ist berechnbar aus fester Menge G
- Holistische Aggregatsfunktion: f kann nur aus den Grundelementen von X berechnet werden

Aggregatstyp	**Beispiel**
Distributiv	$SUM()$, $COUNT()$, $MAX()$, $MIN()$
Algebraisch	$AVG()$ mit $g_1 := SUM()$ und $g_2 := COUNT()$, $STDDEV()$
Holistisch	$MEDIAN()$, $RANK()$, $PERCENTILE()$

Der Würfel

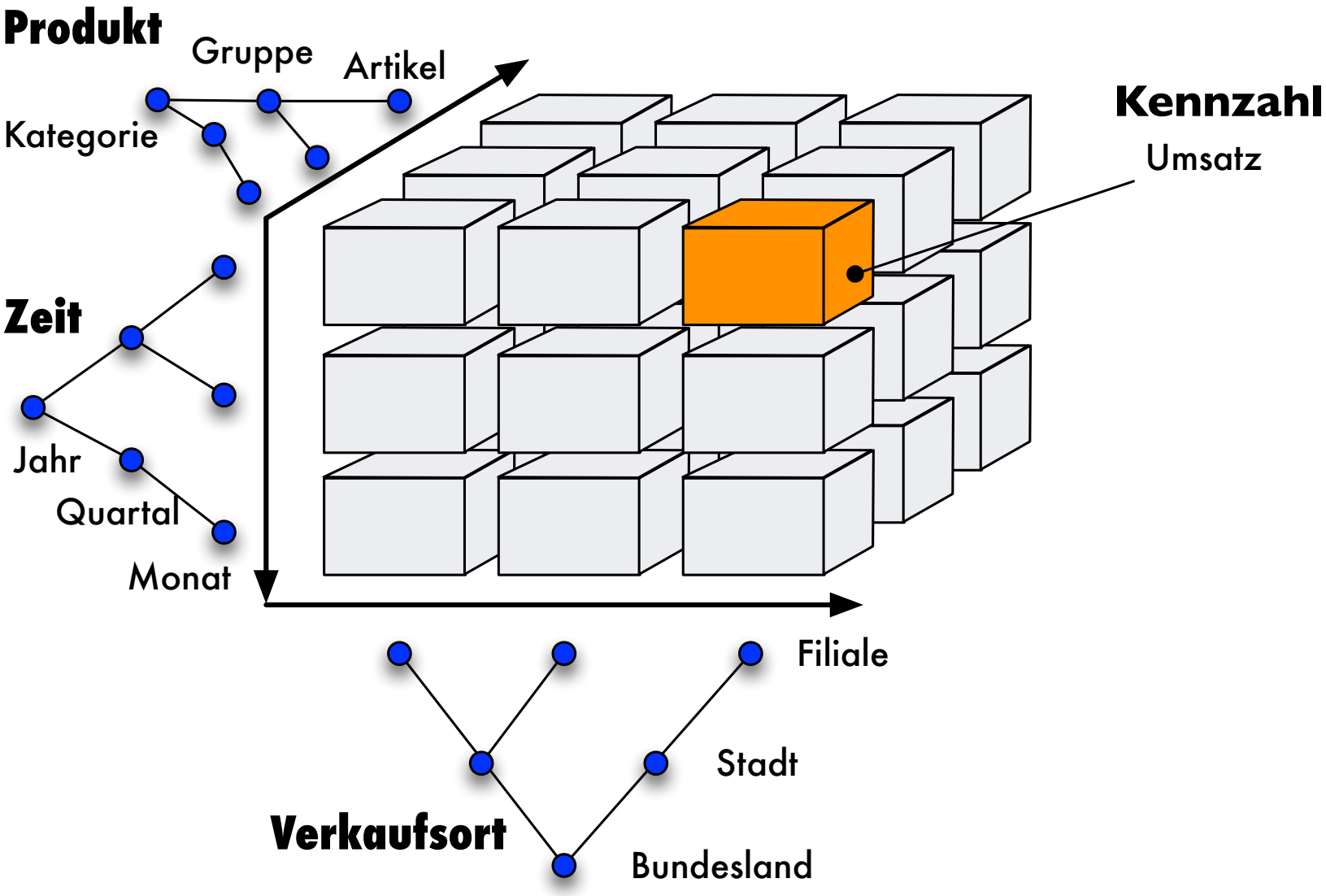
- Würfel (engl. cube, eigentlich Quader): Grundlage der multidimensionalen Analyse
- Kanten $\rightarrow$ Dimensionen
- Zellen $\rightarrow$ ein oder mehrere Kennzahlen (als Funktion der Dimensionen)
- Anzahl der Dimensionen $\rightarrow$ Dimensionalität
- Visualisierung
  - 2 Dimensionen: Tabelle
  - 3 Dimensionen: Würfel
  - > 3 Dimensionen: Multidimensionale Domänenstruktur
- Schema C eines Würfels
  - Menge der Dimensionen(-schemata) DS
  - Menge der Kennzahlen M
- $C = (DS, M) = (\{D^1, \ldots, D^n\}, \{M^1, \ldots, M^m\})$

Orthogonalität

- In multidimensionalen Schemata gilt Orthogonalität, d.h.
  - Keine funktionalen Abhängigkeiten zwischen Attributen unterschiedlicher Dimensionen

$$\forall i, 1 \leq i \leq n, \forall j, 1 \leq j \leq n, i \neq j \neg \exists k, l : D^i.D_k \rightarrow D^j.D_l$$

Multidimensionaler Datenwürfel



Konzeptuelle Modellierung

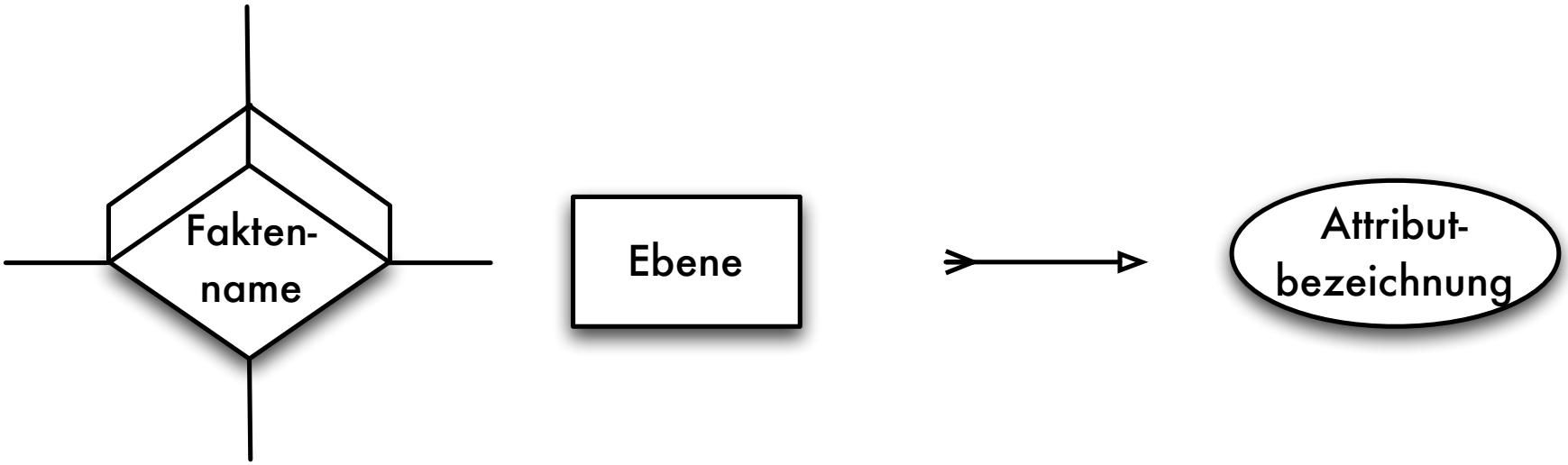
Formale Beschreibung des Fachproblems und der im Anwendungsbereich benötigten Informationsstrukturen

- Probleme konventioneller Entwurfstechniken (ER, UML):
  - Unzureichende Semantik für multidimensionales Datenmodell
  - Hier: Verzicht auf universelle Anwendbarkeit, stattdessen Konzentration auf Analyse
  - Beispiel: Klassifikationsstufe, Fakt $\rightarrow$ Entity ?

ME/R-Modell

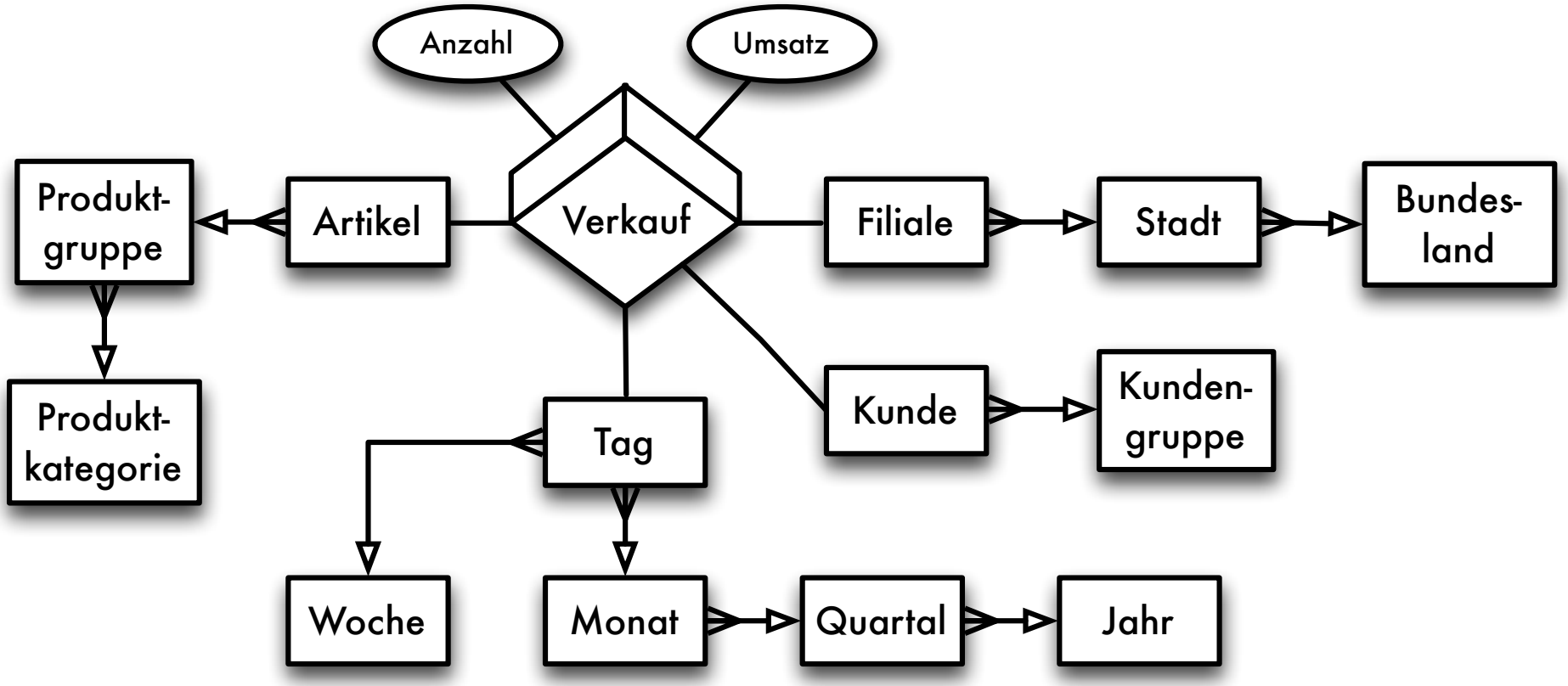
- Multidimensional Entity/Relationship Model
  [Sapia et. al. (1998)]
- Erweiterung des klassischen ER-Modells
  - Entity-Menge „Dimension Level“ (Klassifikationsstufe)
    - ★ Keine explizite Modellierung von Dimensionen
  - n-äre Beziehungsmenge „Fact“
    - ★ Kennzahlen als Attribute der Beziehung
  - Binäre Beziehungsmenge „Classification“ bzw. „Roll-Up“ (Verbindung von Klassifikationsstufen)
    - ★ Definiert gerichteten, nicht-zyklischen Graphen

ME/R: Notationen



Faktenbeziehung | Klassifikationsstufe | Klassifikations-beziehung | Attribut

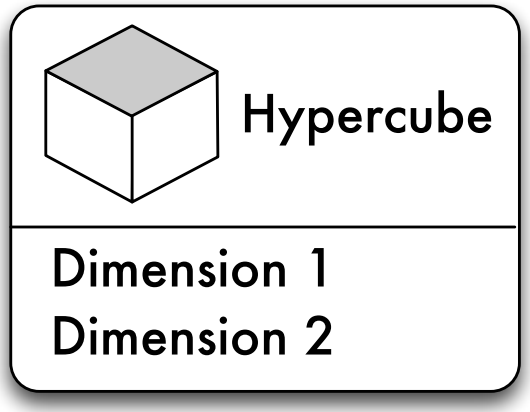
ME/R: Beispiel



ADAPT

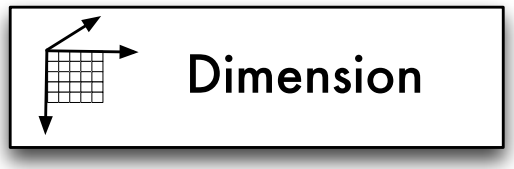
- Application Design for Analytical Processing Technologies
- Bulos 1998 & Bulos 2006
- Berücksichtigt DDL und DML Aspekte
- Fokus auf Dimensionen
- Regeln für Kennzahlen möglich
- Vermischung von Metadaten und Ausprägungen

ADAPT

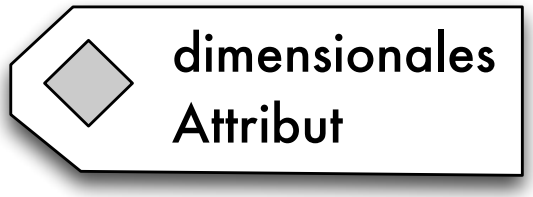

Hypercube
Dimension 1
Dimension 2


f()
Berechnungs-
formel


Hierarchie


Dimension

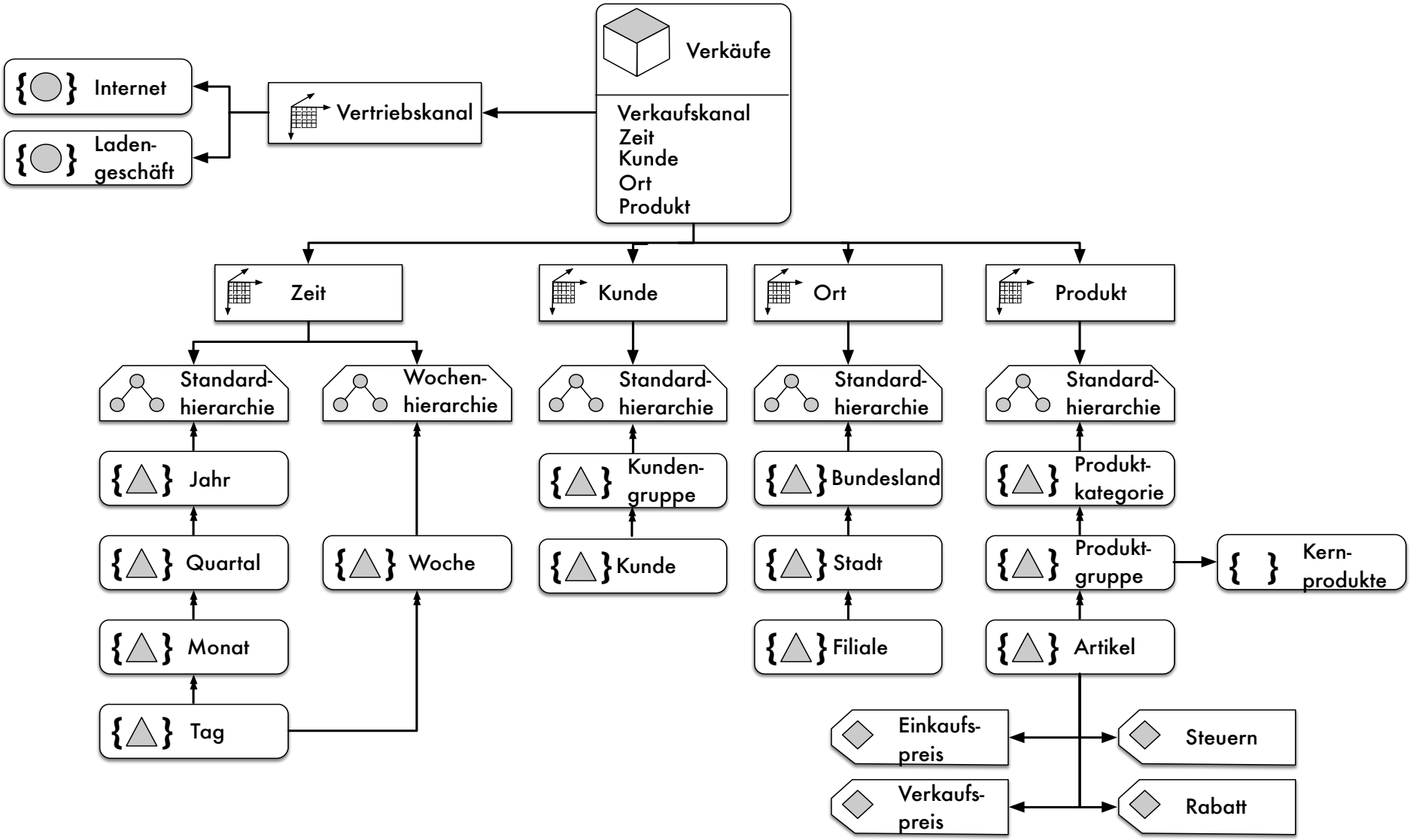

Hierarchie-
stufe


dimensionales
Attribut


Dimensions-
ausprägung


Dimensions-
ausschnitt

ADAPT: Beispiel

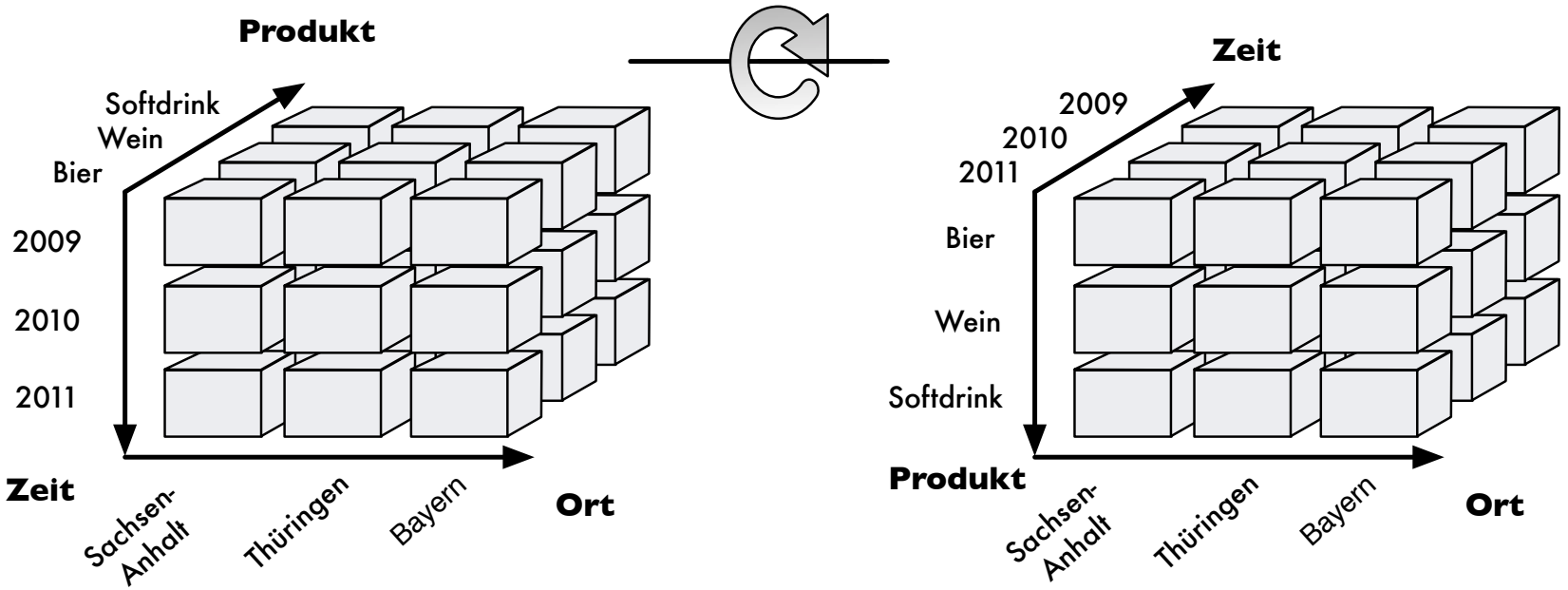


Operationen zur Datenanalyse

- OLAP-Operationen auf multidimensionalen Datenstrukturen
- Standardoperationen
  - Pivotierung
  - Roll-Up, Drill-Down
  - Drill-Across
  - Slice und Dice

Pivotierung / Rotation

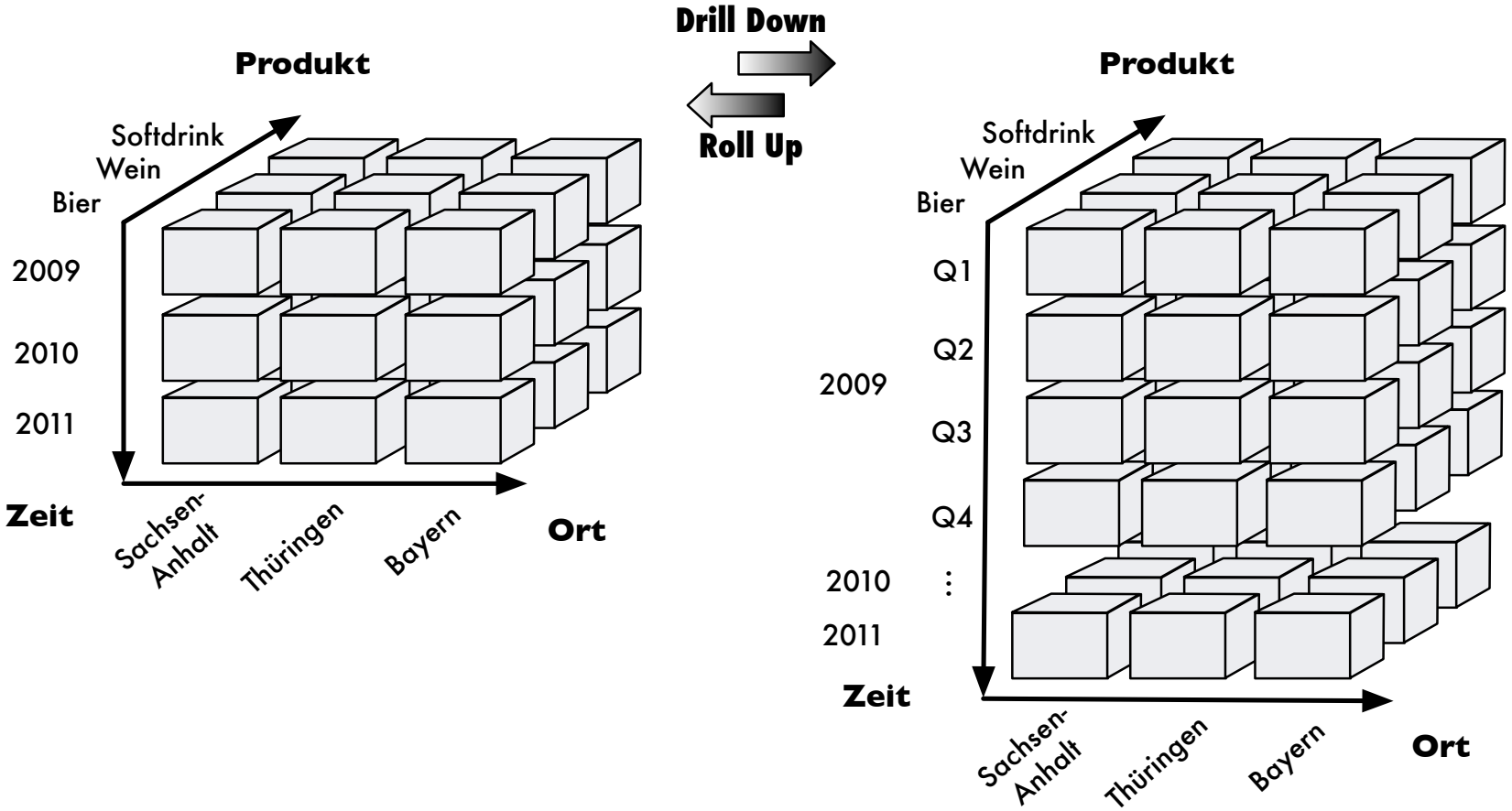
- Drehen des Würfels durch Vertauschen der Dimensionen
- Analyse der Daten aus verschiedenen Perspektiven



Roll-Up, Drill-Down, Drill-Across

- Roll-Up:
  - Erzeugen neuer Informationen durch Aggregierung der Daten entlang des Konsolidierungspfades
  - Dimensionalität bleibt erhalten
  - Beispiel: Tag $\rightarrow$ Monat $\rightarrow$ Quartal $\rightarrow$ Jahr
- Drill-Down:
  - Komplementär zu Roll-Up
  - Navigation von aggregierten Daten zu Detail-Daten entlang der Klassifikationshierarchie
- Drill-Across:
  - Wechsel von einem Würfel zu einem anderen

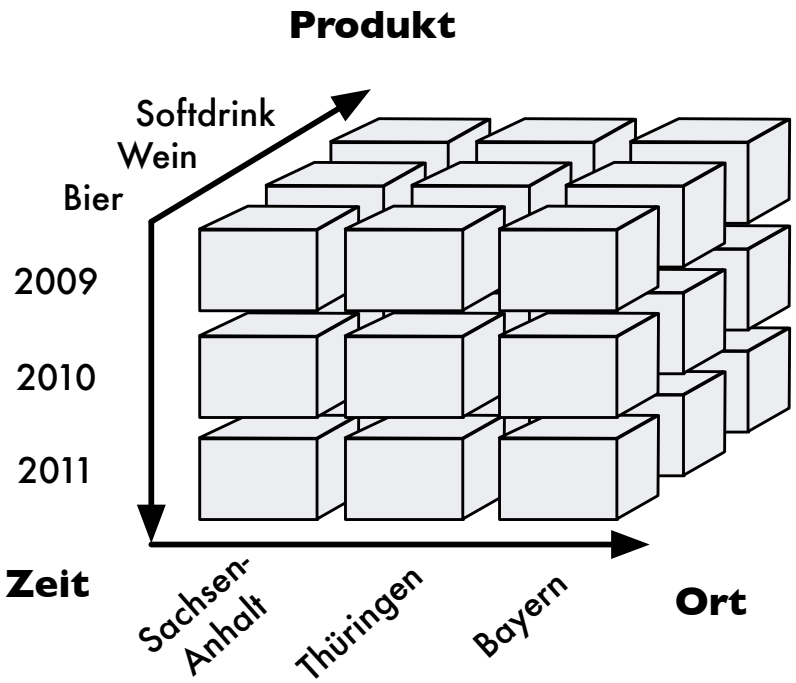
Roll-Up und Drill-Down

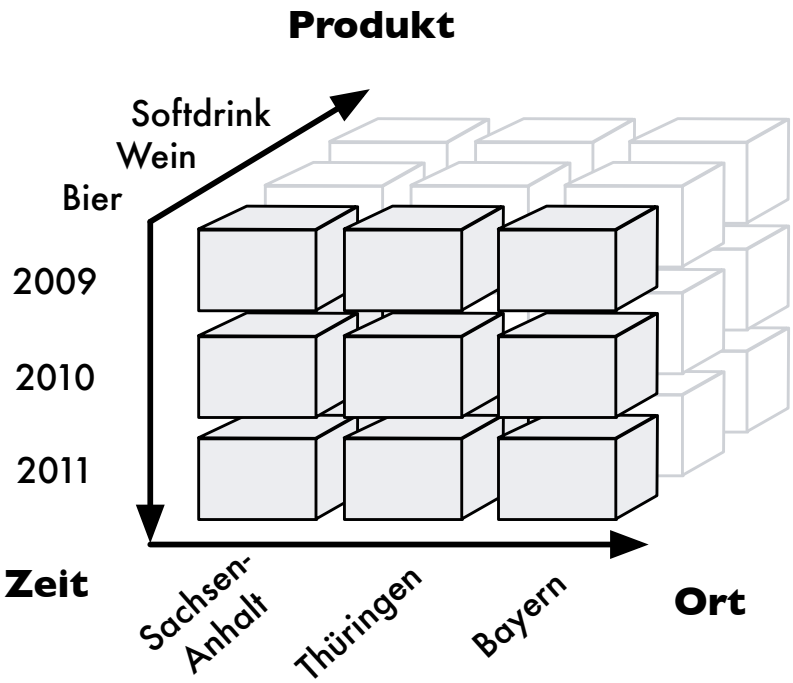


Slice und Dice

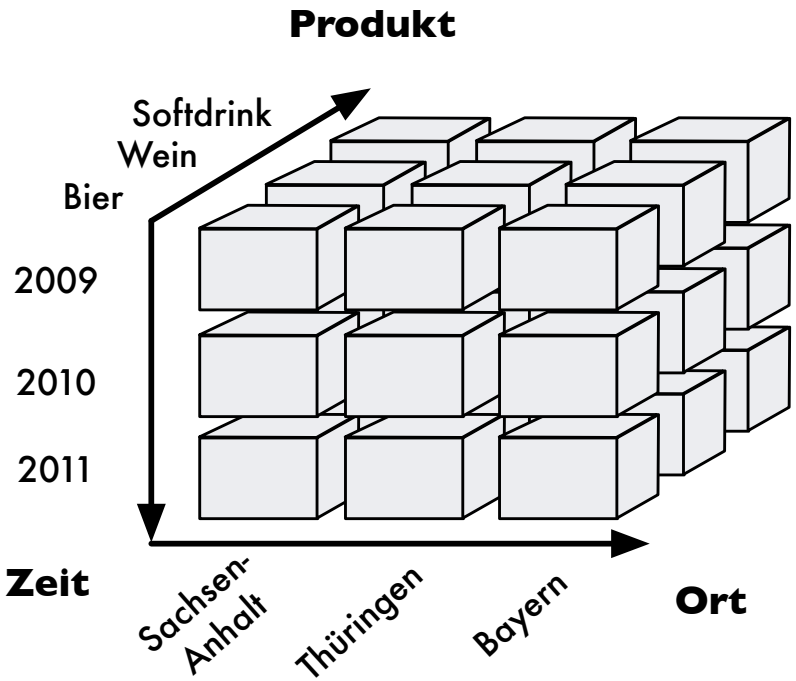
- Erzeugen individueller Sichten
- Slice:
  - Herausschneiden von „Scheiben“ aus dem Würfel
  - Verringerung der Dimensionalität durch Konditionierung der Dimensionen
  - Beispiel: alle Werte des aktuellen Jahres
  - Entspricht der relationalen Selektion in den Dimensionen
- Dice:
  - Herausschneiden einen „Teilwürfels“
  - Erhaltung der Dimensionalität, Veränderung der Hierarchieobjekte
  - Beispiel: die Werte bestimmter Produkte oder Regionen
  - Entspricht der relationalen Selektion mehrerer Dimensionen

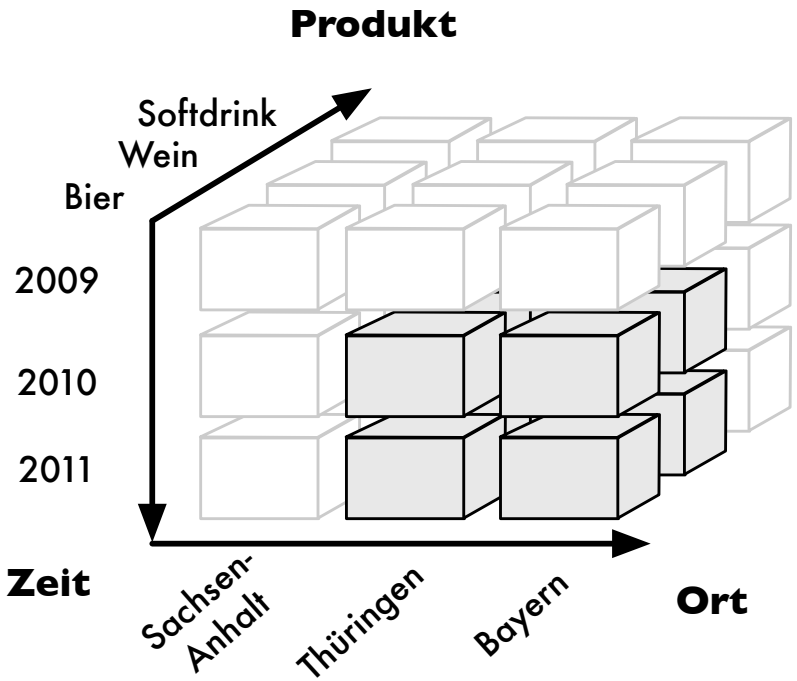
Slice


Produkt
Softdrink
Wein
Bier
2009
2010
2011
Zeit
Sachsen-Anhalt
Thüringen
Bayern
Ort


Produkt
Softdrink
Wein
Bier
2009
2010
2011
Zeit
Sachsen-Anhalt
Thüringen
Bayern
Ort

Dice


Produkt
Softdrink
Wein
Bier
2009
2010
2011
Zeit
Sachsen-Anhalt
Thüringen
Bayern
Ort


Produkt
Softdrink
Wein
Bier
2009
2010
2011
Zeit
Sachsen-Anhalt
Thüringen
Bayern
Ort

Problem des Dimensionswechsel

- Austausch oder Weglassen einer Dimension kann Bestandteil der Analyse sein
- Beispiel: statt Cube by Produkt, Ort und Zeit kann Kunde, Ort und Zeit interessant sein.

Produkt	Filiale	Tag	Verk.
Rotwein	Magdeburg	18.02.10	145
Weißbier	Magdeburg	18.02.10	267
Rotwein	Ilmenau	18.02.10	70
...			

Kundengruppe	Filiale	Tag	Verk.
Weintrinker	Magdeburg	18.02.10	210
Bayer	Magdeburg	18.02.10	407
Genießer	Ilmenau	18.02.10	35
...			

Was bedeutet ein Wechsel der Dimension?

Marginalisierung

- Randsummenbildung bei Entfernen einer Dimension
- Entspricht der relationalen Projektion
- Abhängig vom Typen der Aggregatfunktion $\rightarrow$ Interpretationsproblem

	Anteil Bierflaschen	
	Magdeburg	Ilmenau
1. Halbwoche	95%	88%
2. Halbwoche	68%	54%
GesamtWoche	74%	80%

Umsetzung des multidimensionalen Datenmodells I

- Multidimensionale Sicht
  - Modellierung der Daten
  - Anfrageformulierung
- Interne Verwaltung der Daten erfordert Umsetzung auf
  - Relationale Strukturen (Tabellen)
    $\rightarrow$ ROLAP (relationales OLAP)
    - ★ Verfügbarkeit, Reife der Systeme
  - Multidimensionale Strukturen (direkte Speicherung)
    $\rightarrow$ MOLAP (multidimensionales OLAP)
    - ★ Wegfall der Transformation bzw. Vorwegberechnung
    - ★ Mehrdimensionale Arrays (Fakten) und assoziierte Dimensionslisten
  - Hybride Struktur (Mischform)
    $\rightarrow$ HOLAP (hybrides OLAP)
    - ★ Detaildaten relational abgespeichert (ROLAP)
    - ★ Aggregate werden multidimensional abgespeichert (MOLAP)

Umsetzung des multidimensionalen Datenmodells II

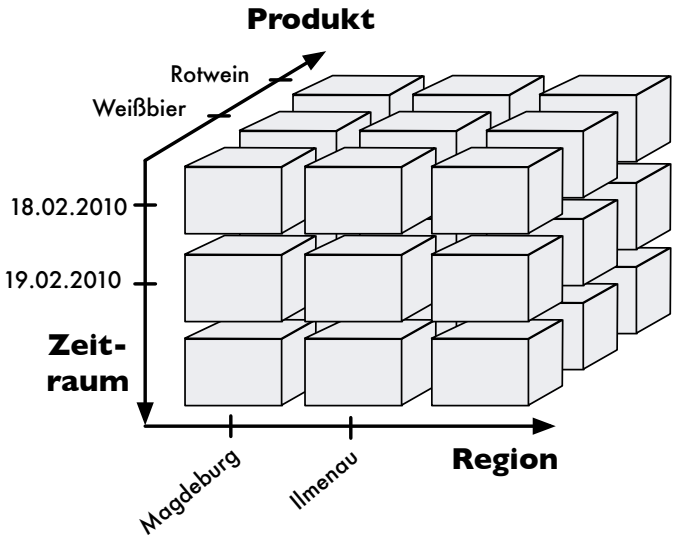
- Aspekte
  - Speicherung
  - Anfrageformulierung bzw. -ausführung

Relationale Speicherung

- Vermeidung des Verlustes anwendungs-bezogener Semantik (aus dem multidimensionalen Modell, z.B. Klassifikationshierarchien)
- Effiziente Übersetzung multidimensionaler Anfragen
- Effiziente Verarbeitung der übersetzten Anfragen
- Einfache Pflege der entstandenen Relationen (z.B. Laden neuer Daten)
- Berücksichtigung der Anfragecharakteristik und des Datenvolumens von Analyseanwendungen

Relationale Umsetzung: Faktentabelle

- Ausgangspunkt: Umsetzung des Datenwürfels ohne Klassifikationshierarchien
  - Dimensionen, Kennzahlen $\rightarrow$ Spalten der Relation
  - Zelle $\rightarrow$ Tupel

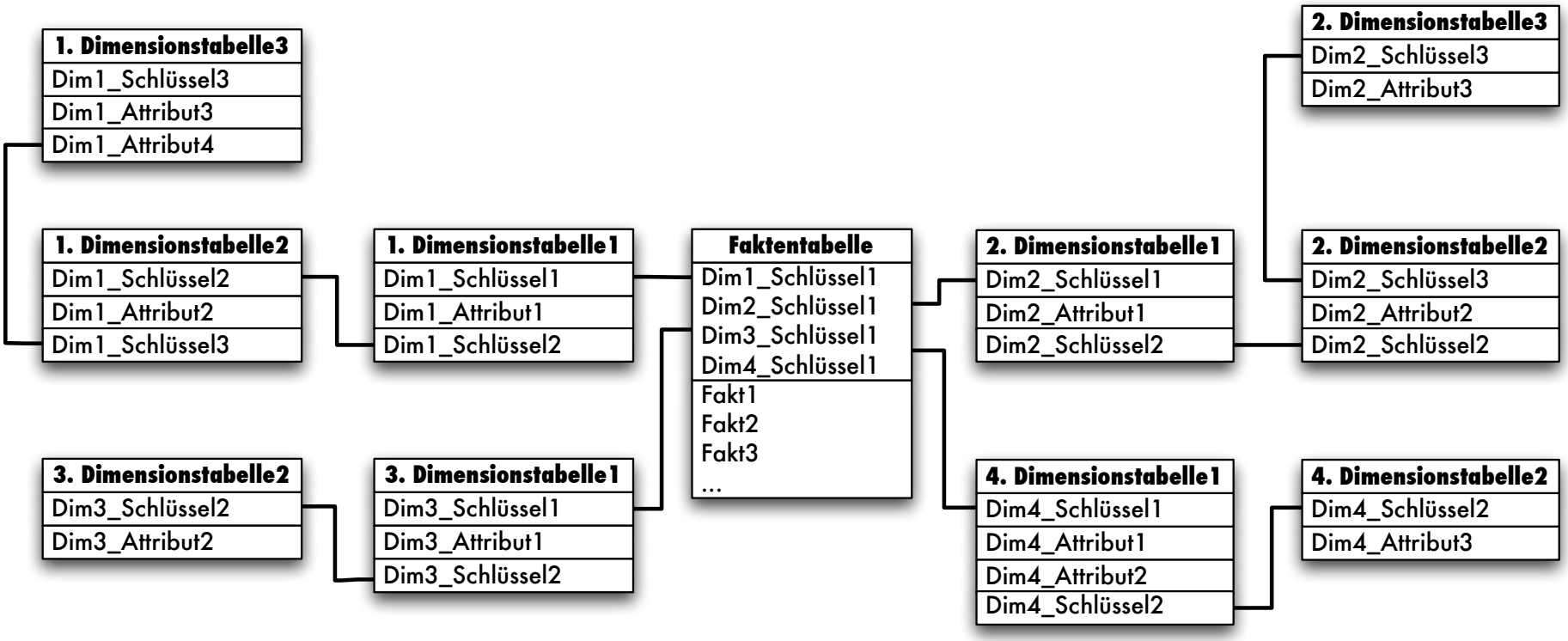


Produkt	Filiale	Tag	Verk.
Rotwein	Magdeburg	18.02.10	145
Weißbier	Magdeburg	18.02.10	267
Rotwein	Ilmenau	18.02.10	70
	...		

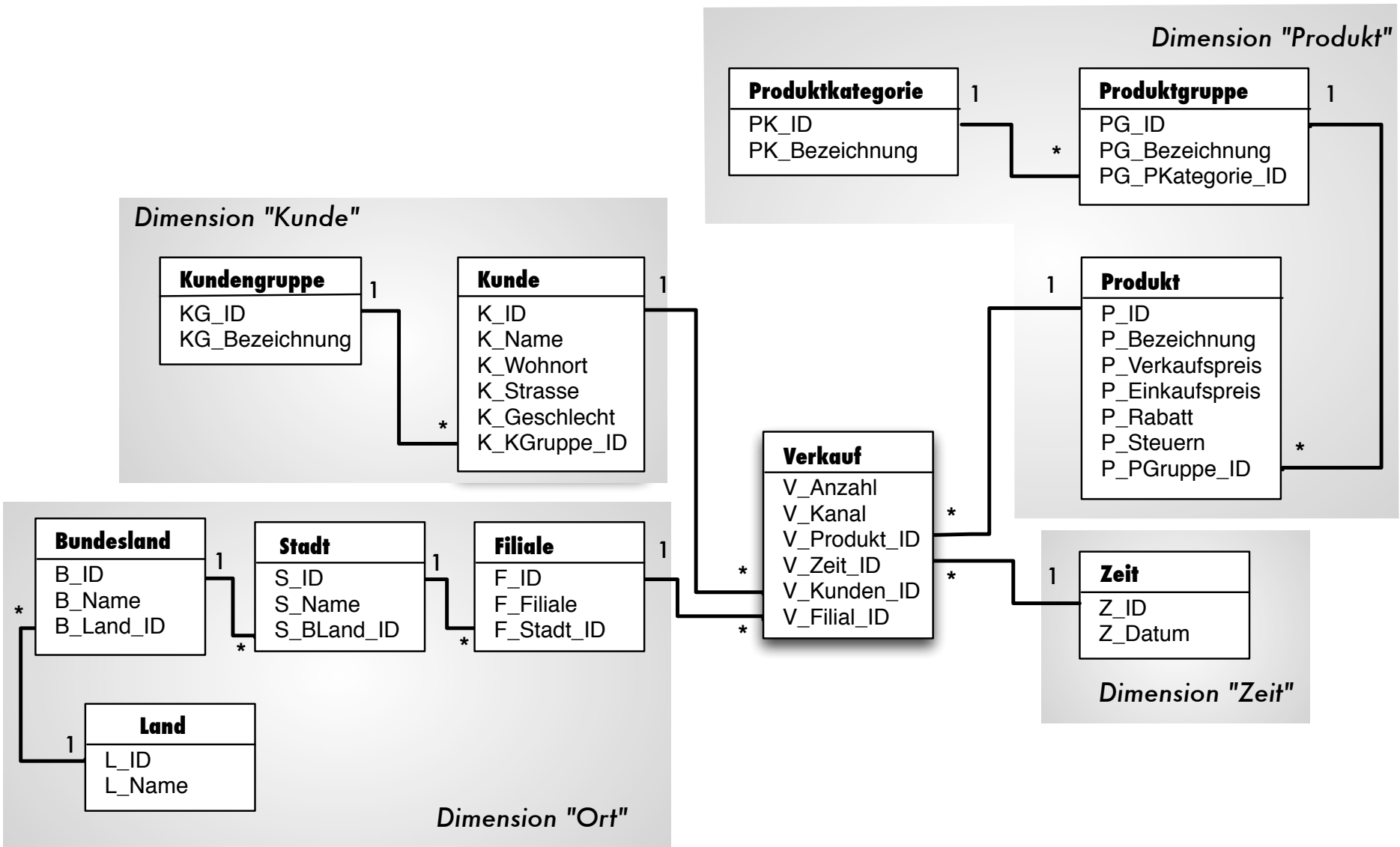
Snowflake-Schema

- Abbildung von Klassifikationen: eigene Tabelle für jede Klassifikationsstufe (z.B. Artikel, Produktgruppe, etc.)
- Dimensionstabelle enthält
  - ID für Klassifikationsknoten
  - Beschreibendes Attribut (z.B. Marke, Hersteller, Bezeichnung)
  - Fremdschlüssel der direkt übergeordneten Klassifikationsstufe
- Faktentabelle enthält (neben Kenngrößen):
  - Fremdschlüssel der jeweils niedrigsten Klassifikationsstufe
  - Fremdschlüssel bilden zusammengesetzte Primärschlüssel für Faktentabelle

Snowflake-Schema: Muster



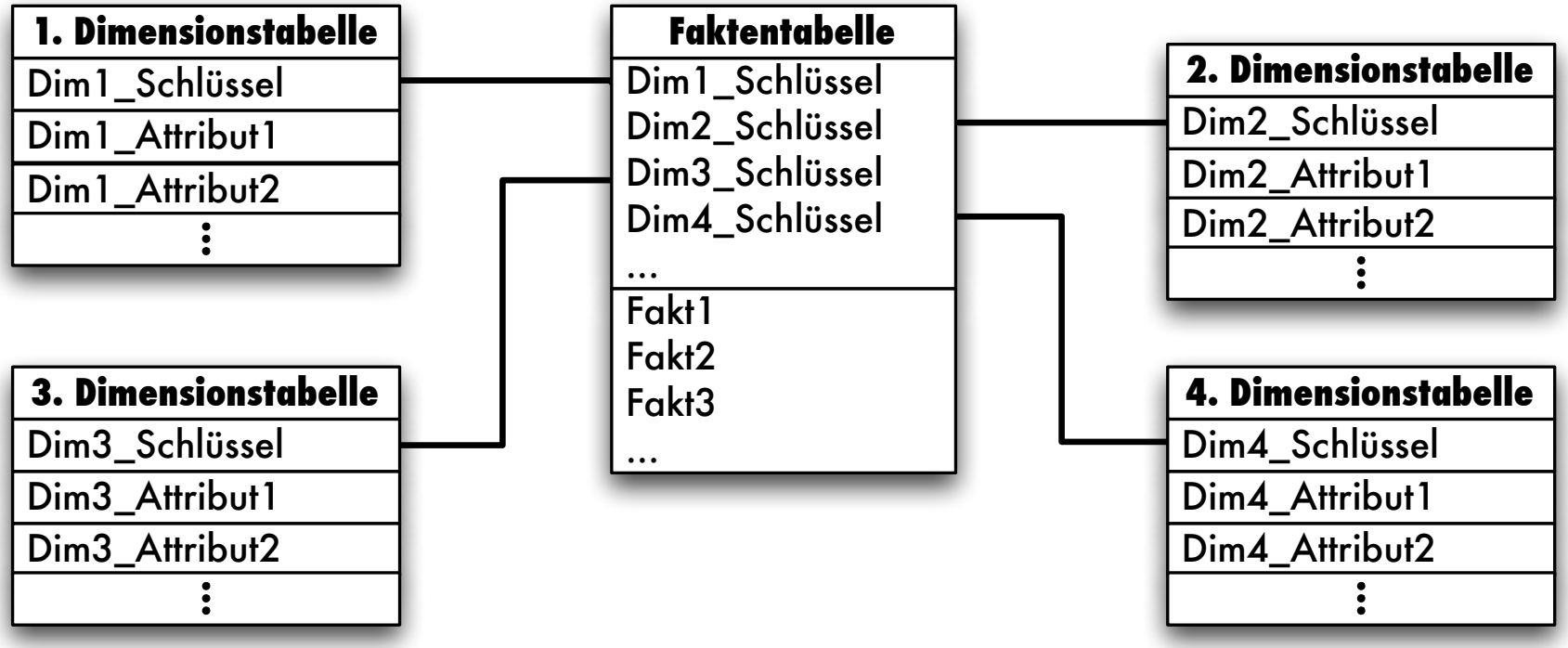
Snowflake-Schema: Beispiel



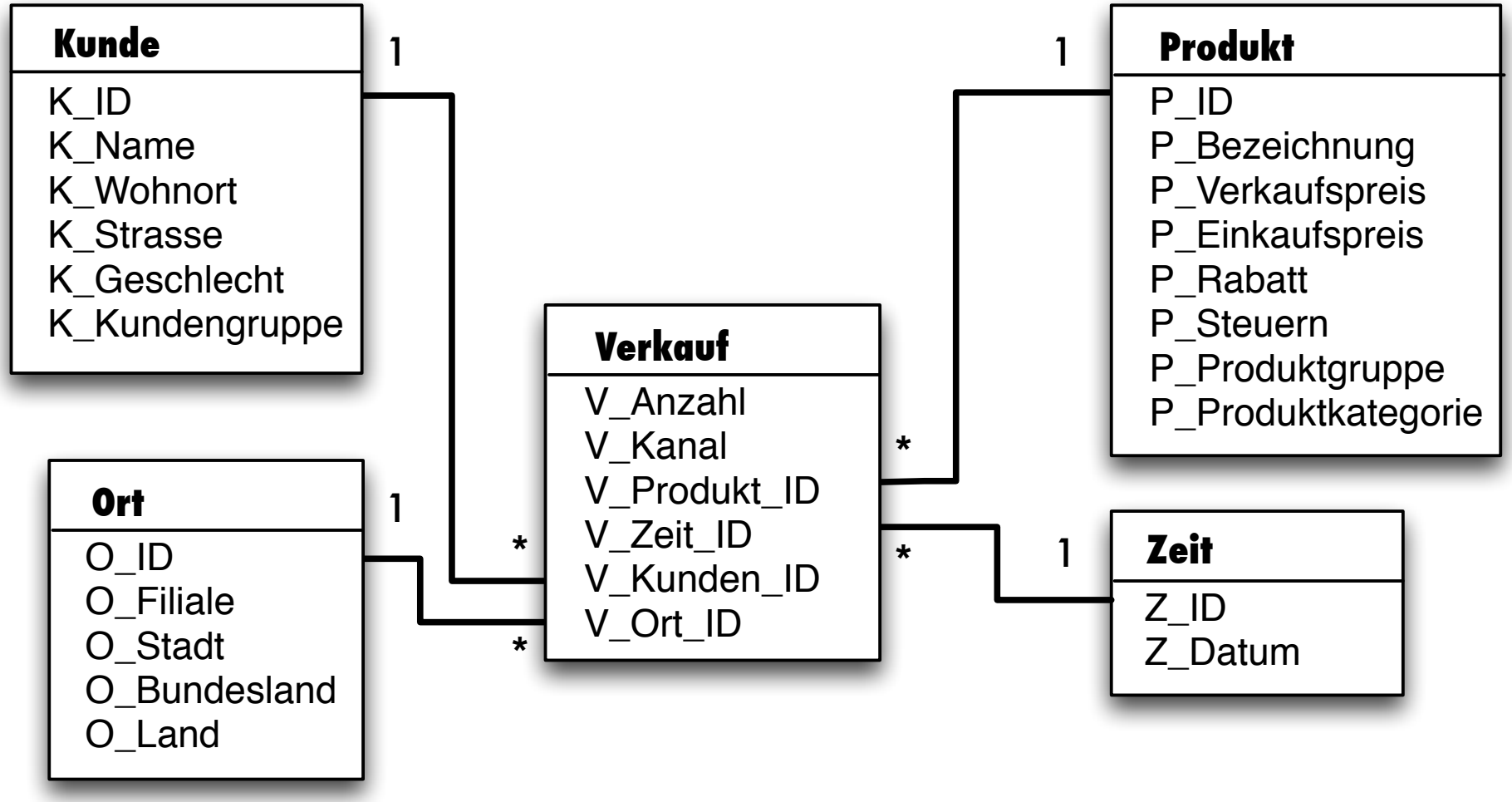
Star-Schema

- Snowflake-Schema ist normalisiert: Vermeidung von Update-Anomalien $\rightarrow$ 3. NF
  - Aber: erfordert Join über mehrere Tabellen!
- Star-Schema:
  - Denormalisierung der zu einer Dimension gehörenden Tabellen $\rightarrow$ 1. NF
  - Für jede Dimension genau eine Dimensionstabelle
  - Redundanzen in der Dimensionstabelle für schnellere Anfragebearbeitung
  - Beispiel: Artikel, Produkt, Produktgruppe etc. als Spalten in einer Tabelle Produkt

Star-Schema: Muster



Star-Schema: Beispiel



Star-Schema formal

- Multidimensionales Schema mit n Dimensionen
  - Dimensionstabellen $D^1, \ldots, D^n$ der Form $D^i(Dim_i\_Key, A_{i,1}, \ldots, A_{i,k_i})$
  - Faktentabelle $F(Dim_1\_Key, \ldots, Dim_n\_Key, f_1, \ldots, f_m)$ mit m Fakten
- Jeder Teil des kompositen Primärschlüssels der Faktentabelle ist Fremdschlüssel zum Primärschlüsselattribut der korrespondierenden Dimension

`CREATE DIMENSION` in Oracle

- Fremdschlüsselbedingungen in SQL ausdrückbar
- Aber: funktionale Beziehungen zwischen Attributen innerhalb einer Dimension nicht spezifizierbar
- Oracle-Erweiterung: `CREATE DIMENSION`
  - „informative“ Zusicherung
  - Korrektheit wird vom DBS nicht überprüft
  - Nutzung beim Query Rewriting über materialisierten Sichten

Beispiel CREATE DIMENSION

```
CREATE DIMENSION ProduktDimension
   LEVEL Artikel IS (Produkt.P_ID)
   LEVEL Produktgruppe IS (Produkt.P_PGruppe_ID)
   LEVEL Produktkategorie IS
      (Produkt.P_Produktkategorie)
HIERARCHY ProduktHierarchie (
   Artikel CHILD OF
   Produktgruppe CHILD OF
   Produktkategorie)
ATTRIBUTE Artikel DETERMINES (P_Bezeichnung,
   P_Verkaufspreis, P_Einkaufspreis,
   P_Rabatt, P_Steuern)
ATTRIBUTE Produktgruppe DETERMINES
   (P_Produktgruppe)
```

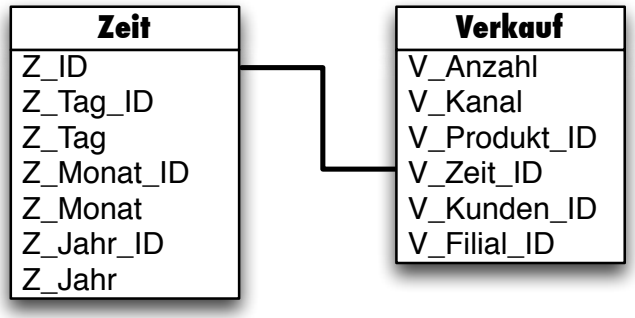
Schlüsselworte

- `LEVEL`
  - Definiert Klassifikationsstufen
- `HIERARCHY`
  - Festlegung der Abhängigkeiten der Klassifikationsstufen
- `ATTRIBUTE ... DETERMINES`
  - Definiert Abhängigkeit zwischen Klassifikationsattribut und dimensionalen Attributen

Snowflake mit `CREATE DIMENSION`


```
CREATE DIMENSION ProduktDimension
   LEVEL Artikel IS (Produkt.P_ID)
   LEVEL Produktgruppe IS (Produktgruppe.PG_ID)
HIERARCHY ProduktHierarchie (
   Artikel CHILD OF Produktgruppe)
JOIN KEY (Produkt.P_PGruppe_ID)
   REFERENCES Produktgruppe
```

Star mit `CREATE DIMENSION`




```
CREATE DIMENSION ZeitDimension
    LEVEL Tag IS (Zeit.Z_Tag_ID)
   LEVEL Monat IS (Zeit.Z_Monat_ID)
   LEVEL Jahr IS (Zeit.Z_Jahr_ID)
HIERARCHY ZeitHierarchie (
   Tag CHILD OF Monat CHILD OF Jahr)
ATTRIBUTE Tag DETERMINES (Z_Tag)
ATTRIBUTE Monat DETERMINES (Z_Monat)
ATTRIBUTE Jahr DETERMINES (Z_Jahr)
```

Vergleich Star und Snowflake

- Charakteristika von DW-Anwendungen
  - Typischerweise Einschränkungen in Anfragen auf höherer Granularitätsstufe (Join-Operationen)
  - Geringes Datenvolumen der Dimensionstabellen im Vergleich zu Faktentabellen
  - Seltene Änderungen an Klassifikationen (Gefahr von Update-Anomalien)
- Vorteile des Star-Schemas
  - Einfache Struktur (vereinfachte Anfrageformulierung)
  - Einfache und flexible Darstellung von Klassifikationshierarchien (Spalten in Dimensionstabellen)
  - Effiziente Anfrageverarbeitung innerhalb einer Dimension (keine Join-Operation notwendig)

Annahmen für Kostenbetrachtungen

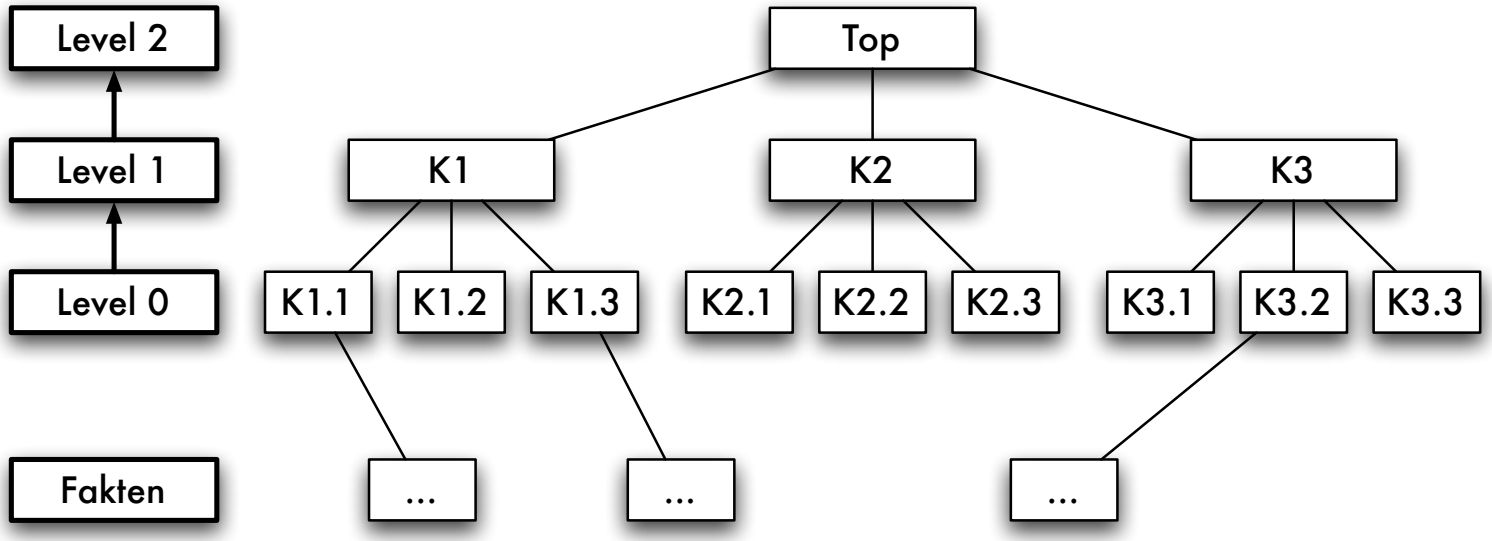
- n Dimensionen (D^n), je K Klassifikationsstufen plus Top
- Jeder Klassifikationsknoten hat 3 Kinder
  - Level $i = K$: $1 = 3^0$ Knoten (Top)
  - Level $i = K - 1$: $3 = 3^1$ Knoten
  - Level $i = K - 2$: $9 = 3^2$ Knoten
  - ...
  - Level $i = 0$: Höchste Granularität, 3^K Knoten
  - $N_D = \sum_{i=0\ldots K}$ Knoten pro Dimension

$$N_D = \sum_{i=0\ldots K} 3^i$$

- M Fakten, gleichverteilt in Dimensionen
- Attribut: b Bytes; Knoten haben nur ID; m Faktattribute

Volle Klassifikationshierarchie

- Zu jedem Knoten gibt es (gleich viele) Fakten



Anfragekosten Star vs. Snowflake

- Speicherplatz Snowflake:

$$(((n+m)\cdot M)+n\cdot N_D)\cdot b$$

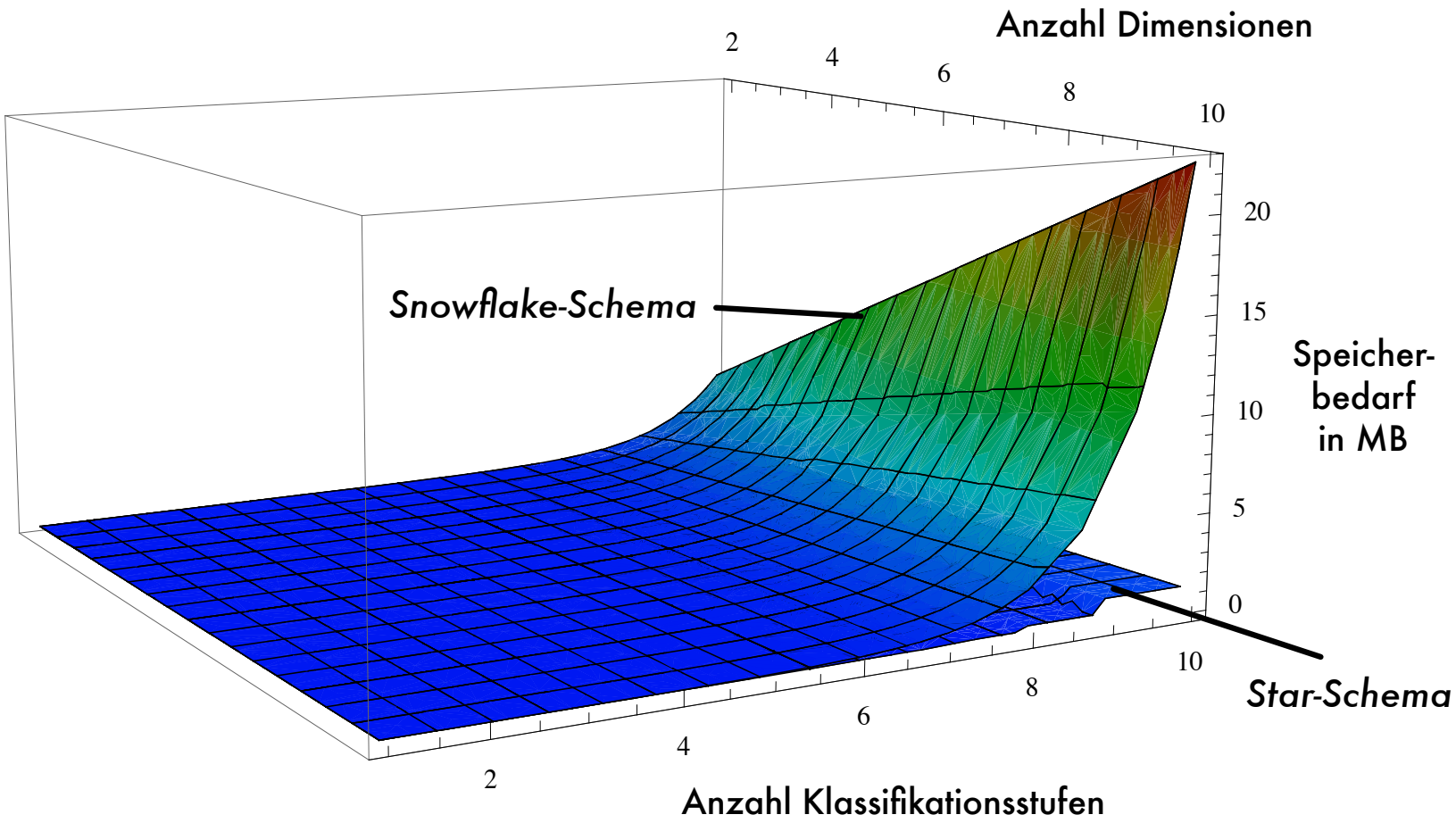
  - n Anzahl Fremdschlüssel in Faktentabelle
  - m Anzahl Faktenattribute
  - $n \cdot N_D$ Ein Tupel pro Klassifikationsknoten

- Speicherplatz Star:

$$(((n+m)\cdot M)+n\cdot 3^K\cdot K)\cdot b$$

  - $n \cdot 3^K$ Ein Tupel pro Klassifikationsknoten Level 0
  - K Ein Attribut pro Klassifikationsstufe

Kosten Star vs. Snowflake



Anfrage Star und Snowflake

- Anfrage: Verkäufe der Produktgruppe „Wein“ pro Stadt und Jahr
- Snowflake-Schema:

```
SELECT  S_Name, YEAR(Z_Datum), SUM(V_Anzahl)
FROM Verkauf, Filiale, Stadt, Produkt, Produktgruppe, Zeit
WHERE V_Produkt_ID = P_ID AND
   P_PGruppe_ID = PG_ID AND
   V_Filial_ID = F_ID AND
   F_Stadt_ID = S_ID AND
   V_Zeit_ID = Z_ID AND
   PG_Bezeichnung = 'Wein'
GROUP BY S_Name, YEAR(Z_Datum)
```

- Anzahl der Joins: 5
  (steigt linear mit Anzahl der Aggregationspfade)

Anfrage Star und Snowflake (2)

- Anfrage für Star-Schema:

```
SELECT O_Stadt, YEAR(Z_Datum), SUM(V_Anzahl)
FROM Verkauf, Ort, Produkt, Zeit
WHERE V_Produkt_ID = P_ID AND
   V_Zeit_ID = Z_ID AND
   V_Ort_ID = O_ID AND
   P_Produktgruppe = 'Wein'
GROUP BY O_Stadt, YEAR(Z_Datum)
```

- Anzahl der Joins: 3
  (unabhängig von der Länge der Aggregationspfade)

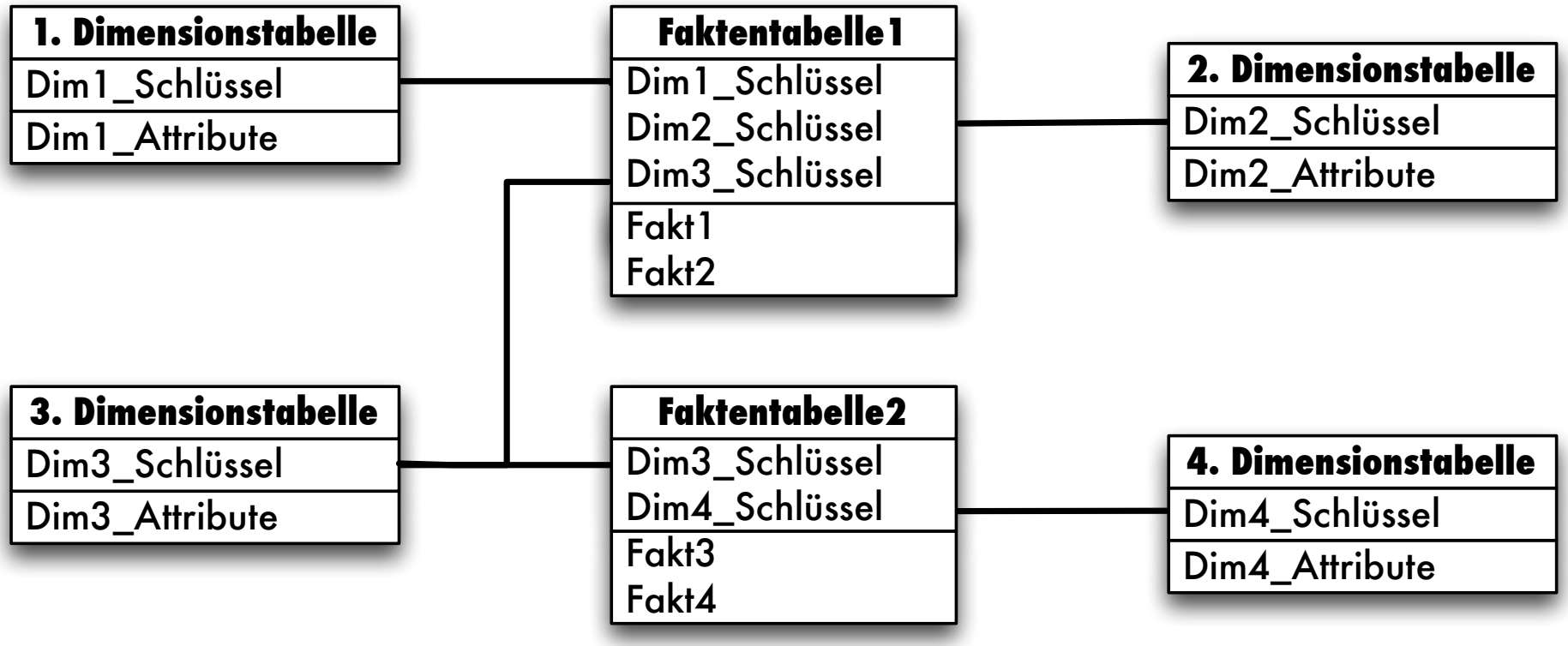
Mischformen

- Abbildung einzelner Dimensionen analog Snowflake-Schema oder Star-Schema
- Entscheidungskriterien:
  - Änderungshäufigkeit der Dimensionen:
    - ★ Reduzierung des Pflegeaufwandes durch Normalisierung (Snowflake)
  - Anzahl der Klassifikationsstufen einer Dimension:
    - ★ Mehr Klassifikationsstufen $\rightarrow$ größere Redundanz im Star-Schema
  - Anzahl der Dimensionselemente:
    - ★ Einsparung durch Normalisierung bei vielen Elementen einer Dimension auf niedrigster Klassifikationsstufe
  - Materialisierung von Aggregaten:
    - ★ Performance-Verbesserung durch Normalisierung bei materialisierten Aggregaten für eine Klassifikationsstufe

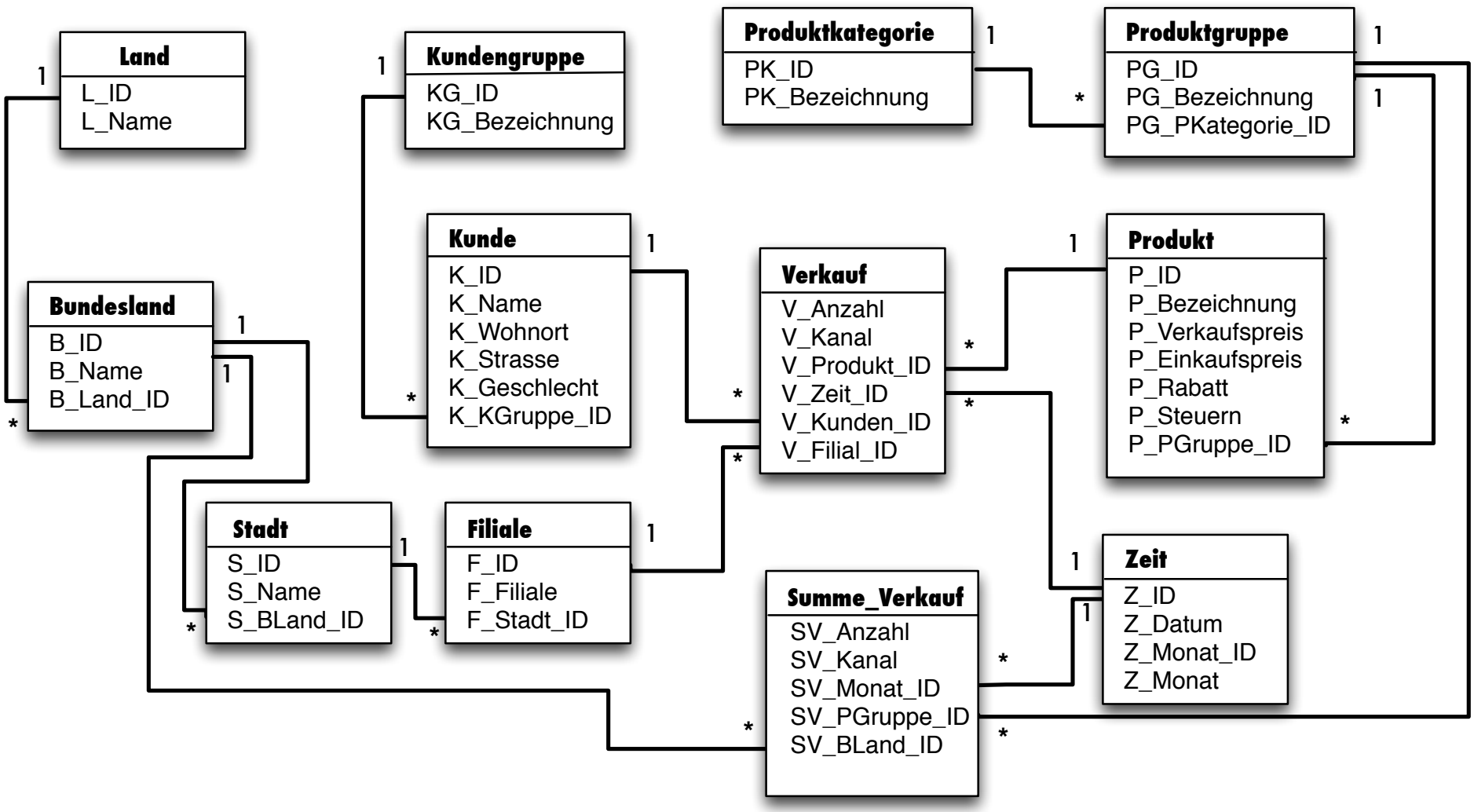
Galaxie-Schema

- Star- und Snowflakeschema
  - Eine Faktentabelle
  - Mehrere Kennzahlen nur möglich bei gleichen Dimensionen
- Galaxie-Schema
  - Mehrere Faktentabellen
  - Teilweise mit gleichen Dimensionstabellen verknüpft
  - Auch: Multi-Faktentabellen-Schema, Multi-Cube, Hyper-Cube

Galaxie-Schema: Muster



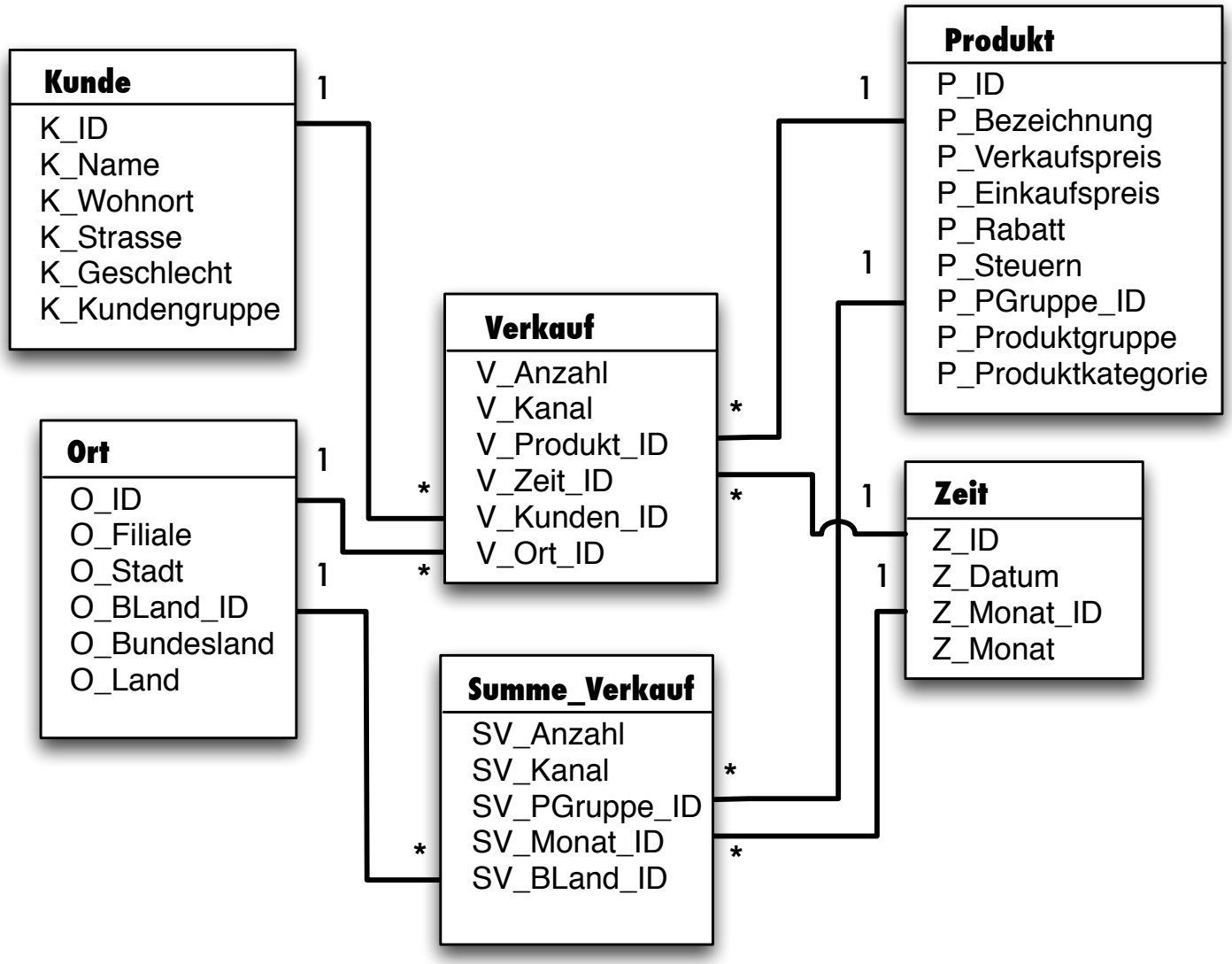
Galaxie-Schema: Beispiel



Fact Constellation

- Speicherung vorberechneter Aggregate in Faktentabelle
  - Beispiel: Umsatz für Region
  - Unterscheidung in Dimensionstabelle über spezielle Attribute (Bsp.: „Stufe“)
- Alternative: Auslagerung in eigene Faktentabelle
  - Fact-Constellation-Schema (Spezialfall eines Galaxie-Schemas)

Fact Constellation: Beispiel



Darstellung von Klassifikationshierarchien

- Horizontal: Modellierung der Stufen der Klassifikationshierarchie als Spalten der denormalisierten Dimensionstabelle
  - Vorteil:
    - ★ Einschränkungen auf höherer Granularität ohne Join
  - Nachteile:
    - ★ Duplikateliminierung beim Anfragen bestimmter Stufen (Bsp.: Produktgruppe innerhalb einer Kategorie)
    - ★ Schemaänderung beim Hinzufügen neuer Stufen

Produkt_ID	Artikel	Produktgruppe	Produktkategorie
1234	Immer Ultra	Hygiene	Kosmetik
1235	Putzich	Hygiene	Kosmetik
2345	Rohrfrei	Reiniger	Haushalt

Darstellung von Klassifikationshierarchien

- Vertikal (rekursiv): normalisierte Dimensionstabelle mit Attributen
  - Dimensions_ID: Schlüssel für Faktentabelle
  - Eltern_ID: Attributwert der Dimensions-ID der nächsthöheren Stufe
- Vorteile:
  - Einfache Änderung am Klassifikationsschema
  - Einfache Behandlung vorberechneter Aggregate
- Nachteil:
  - Self-Join für Anfragen einzelner Stufen (Bsp.: Produktgruppe innerhalb einer Kategorie)

Dimensions_ID	Eltern_ID
Immer Ultra	Hygiene
Hygiene	Kosmetik
Putzich	Hygiene

Darstellung von Klassifikationshierarchien

- Kombiniert: Verbindung beider Strategien
  - Repräsentation der Klassifikationsstufen als Spalten (jedoch generische Bezeichnung)
  - Speicherung der Knoten aller höheren Stufen als Tupel
  - Zusätzliches Attribut „Stufe“ $\rightarrow$ Angabe der bezeichneten Klassifikationsstufe

Dimensions_ID	Stufe1_ID	Stufe2_ID	Stufe
Immer Ultra	Hygiene	Kosmetik	0
Putzich	Hygiene	Kosmetik	0
Hygiene	Kosmetik	NULL	1
Kosmetik	NULL	NULL	2

Vermeidung von Semantikverlusten

- Semantikverlust bei relationaler Abbildung:
  - Unterscheidung zwischen Kennzahl und Dimension (Attribute der Faktentabelle)
  - Attribute von Dimensionstabellen (beschreibend, Aufbau der Hierarchie)
  - Aufbau der Dimensionen (Drill-Pfade)
- Ausweg:
  - Erweiterung des Systemkatalogs um Metadaten für multidimensionale Anwendungen
  - Beispiel: CREATE DIMENSION, HIERARCHY in Oracle

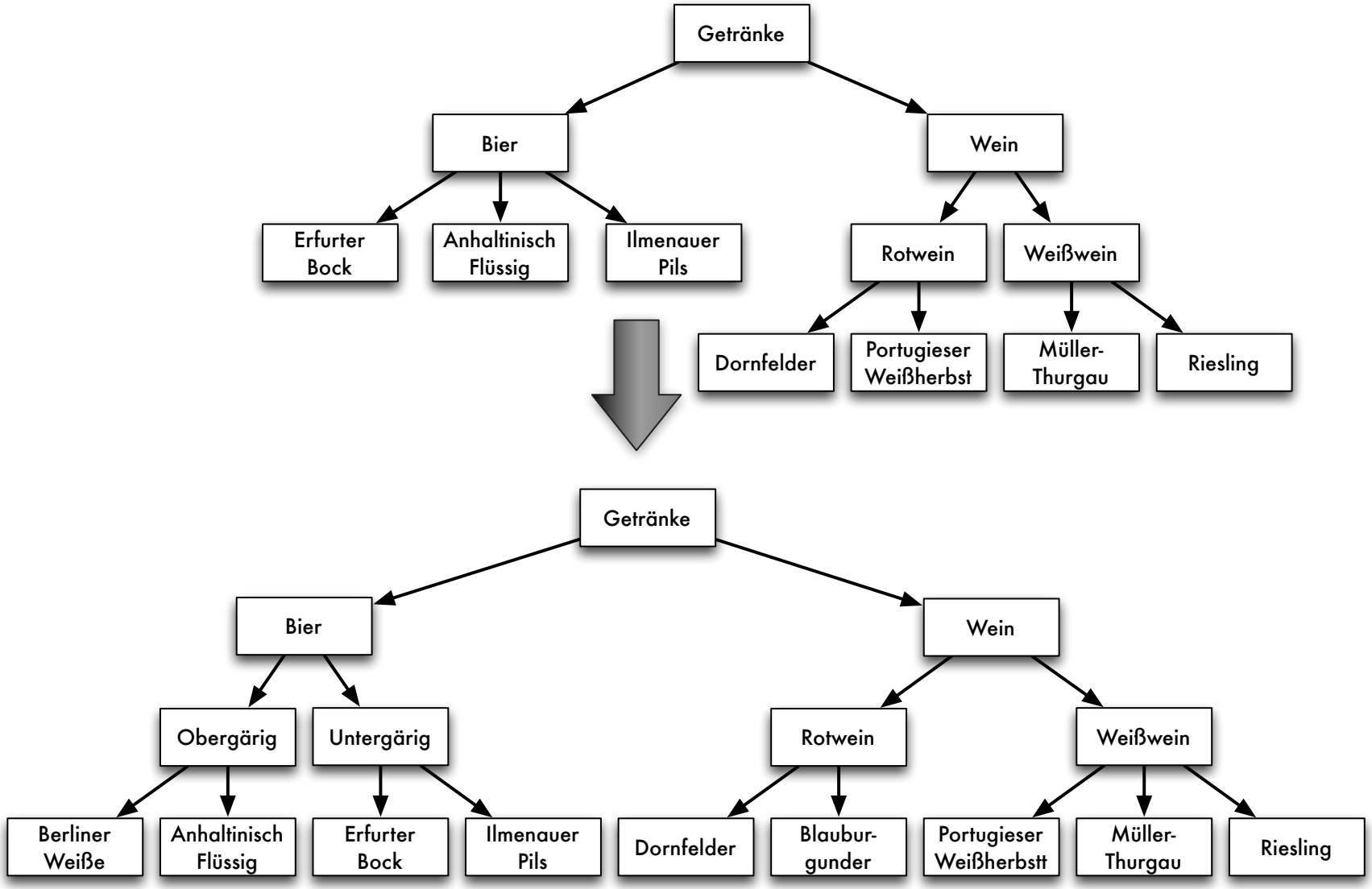
Probleme der relationalen Umsetzung

- Transformation multidimensionaler Anfragen in relationale Repräsentation notwendig $\rightarrow$ komplexe Anfragen
- Einsatz komplexer Anfragewerkzeuge notwendig (OLAP-Werkzeuge)
- Semantikverlust
- Daher: direkte multidimensionale Speicherung $\rightarrow$ siehe Teil VI

Slowly Changing Dimensions

- Historisierung: Änderungen von Attributsausprägungen, Beziehungen und Entitäten im Zeitablauf
- Strukturveränderungen in den Dimensionen
- Schemaveränderungen in den Dimensionen
- Einfluss auf Analysen

Slowly Changing Dimensions: Beispiel



Analyseanforderungen

- nach aktueller Struktur: **as is**
- nach definierter historischer Struktur: **as of**
- gemäß historischer Wahrheit: **as posted**
- **vergleichbare Resultate**

Beispiel für Slowly Changing Dimensions

Produktdimension 2010

Produkt	Produktgruppe
Müller-Thurgau	Weißwein
Riesling	Weißwein
Dornfelder	Rotwein
Portugieser Weißherbst	Rotwein

Produktdimension 2011

Produkt	Produktgruppe
Müller-Thurgau	Weißwein
Riesling	Weißwein
Portugieser Weißherbst	Weißwein
Blauburgunder	Rotwein
Dornfelder	Rotwein

Fakten

Produkt	Jahr	Pakete
Dornfelder	2010	1000
Müller-Thurgau	2010	1000
Portugieser Weißherbst	2010	1000
Riesling	2010	1000
Blauburgunder	2011	1000
Dornfelder	2011	1000
Müller-Thurgau	2011	1000
Portugieser Weißherbst	2011	1000
Riesling	2011	1000

Berichtsanforderungen

aktuelle Struktur

Produktgruppe	Pakete 2010	Pakete 2011
Weißwein	3000	3000
Rotwein	1000	2000

alte Struktur

Produktgruppe	Pakete 2010	Pakete 2011
Weißwein	2000	2000
Rotwein	2000	2000

historische Wahrheit

Produktgruppe	Pakete 2010	Pakete 2011
Weißwein	2000	3000
Rotwein	2000	2000

vergleichbare Resultate

Produktgruppe	Pakete 2010	Pakete 2011
Weißwein	2000	2000
Rotwein	1000	1000

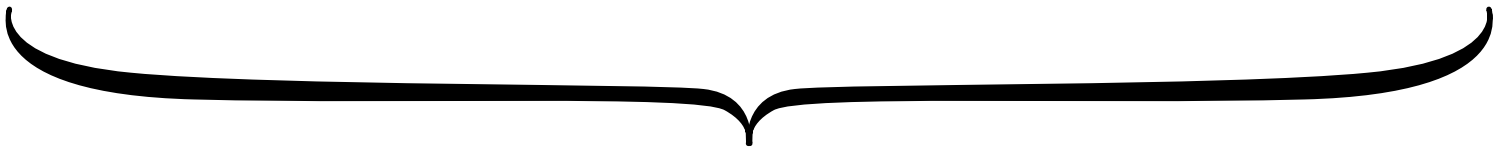
. . . mit aktueller Struktur

Produktdimension 2011

Produkt	Produktgruppe
Müller-Thurgau	Weißwein
Riesling	Weißwein
Portugieser Weißherbst	Weißwein
Blauburgunder	Rotwein
Dornfelder	Rotwein

Fakten

Produkt	Jahr	Pakete
Dornfelder	2010	1000
Müller-Thurgau	2010	1000
Portugieser Weißherbst	2010	1000
Riesling	2010	1000
Blauburgunder	2011	1000
Dornfelder	2011	1000
Müller-Thurgau	2011	1000
Portugieser Weißherbst	2011	1000
Riesling	2011	1000



aktuelle Struktur

Produktgruppe	Pakete 2010	Pakete 2011
Weißwein	3000	3000
Rotwein	1000	2000

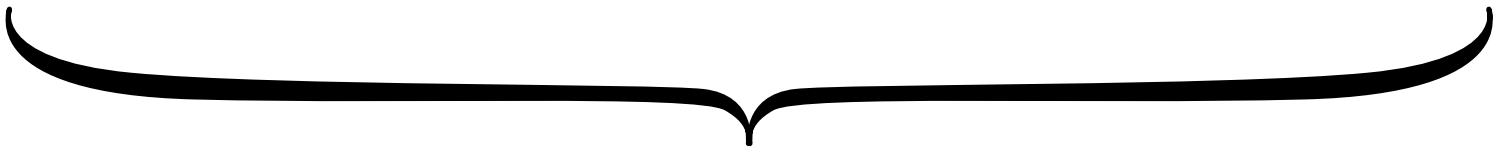
. . . nach alter Struktur

Produktdimension 2010

Produkt	Produktgruppe
Müller-Thurgau	Weißwein
Riesling	Weißwein
Dornfelder	Rotwein
Portugieser Weißherbst	Rotwein

Fakten

Produkt	Jahr	Pakete
Dornfelder	2010	1000
Müller-Thurgau	2010	1000
Portugieser Weißherbst	2010	1000
Riesling	2010	1000
Blauburgunder	2011	1000
Dornfelder	2011	1000
Müller-Thurgau	2011	1000
Portugieser Weißherbst	2011	1000
Riesling	2011	1000



alte Struktur

Produktgruppe	Pakete 2010	Pakete 2011
Weißwein	2000	2000
Rotwein	2000	2000

. . . nach historischer Wahrheit

Produktdimension 2010

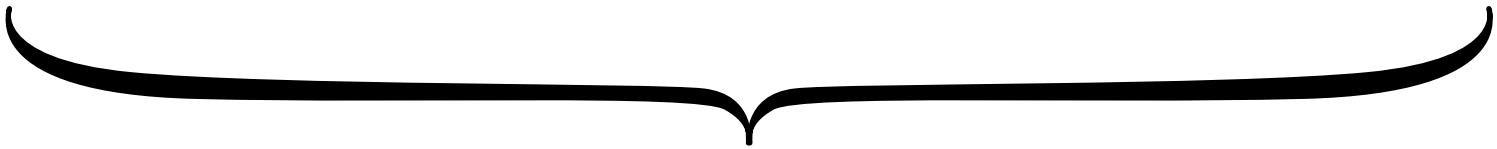
Produkt	Produktgruppe
Müller-Thurgau	Weißwein
Riesling	Weißwein
Dornfelder	Rotwein
Portugieser Weißherbst	Rotwein

Produktdimension 2011

Produkt	Produktgruppe
Müller-Thurgau	Weißwein
Riesling	Weißwein
Portugieser Weißherbst	Weißwein
Blauburgunder	Rotwein
Dornfelder	Rotwein

Fakten

Produkt	Jahr	Pakete
Dornfelder	2010	1000
Müller-Thurgau	2010	1000
Portugieser Weißherbst	2010	1000
Riesling	2010	1000
Blauburgunder	2011	1000
Dornfelder	2011	1000
Müller-Thurgau	2011	1000
Portugieser Weißherbst	2011	1000
Riesling	2011	1000



historische Wahrheit

Produktgruppe	Pakete 2010	Pakete 2011
Weißwein	2000	3000
Rotwein	2000	2000

. . . mit vergleichbaren Resultaten

Produktdimension 2010

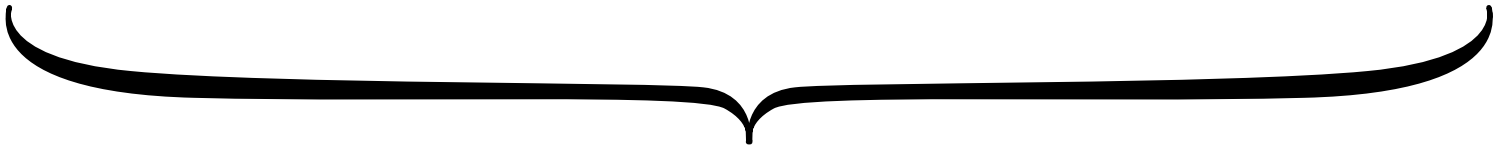
Produkt	Produktgruppe
Müller-Thurgau	Weißwein
Riesling	Weißwein
Dornfelder	Rotwein
Portugieser Weißherbst	Rotwein

Produktdimension 2011

Produkt	Produktgruppe
Müller-Thurgau	Weißwein
Riesling	Weißwein
Portugieser Weißherbst	Weißwein
Blauburgunder	Rotwein
Dornfelder	Rotwein

Fakten

Produkt	Jahr	Pakete
Dornfelder	2010	1000
Müller-Thurgau	2010	1000
Portugieser Weißherbst	2010	1000
Riesling	2010	1000
Blauburgunder	2011	1000
Dornfelder	2011	1000
Müller-Thurgau	2011	1000
Portugieser Weißherbst	2011	1000
Riesling	2011	1000



vergleichbare Resultate

Produktgruppe	Pakete 2010	Pakete 2011
Weißwein	2000	2000
Rotwein	1000	1000

Realisierungen

- Anpassung des historischen Datenmaterials an neue Strukturen
  kleiner Datenbestand, aber alte Strukturen sind nicht vorhanden
- Separate Speicherung des historischen Datenbestandes zusätzlich zu neuen Strukturen
  großer Datenbestand und Komplexität für Anwender, aber alte Strukturen abrufbar
- Aufbau paralleler Hierarchien mit allen Strukturen
  Dimensionsstruktur komplex, beliebiges Reporting möglich
- Gültigkeitsstempel
  beliebige Strukturen abrufbar, aber Performancefrage

„As is“-Szenario

- Einfach zu implemetieren
- Keine Historie
- Nur für „as is“

Getraenk_ID	Produkt	Produktgruppe
1	Dornfelder	Rotwein
2	Müller-Thurgau	Weißwein

Getraenk_ID	Produkt	Produktgruppe
1	Dornfelder	Rotwein
2	Müller-Thurgau	Weißwein
~~3~~	~~Portugieser Weißherbst~~	~~Rotwein~~
3	Portugieser Weißherbst	Weißwein

„As is“ & quasi „as of“-Szenario

- Etwas höherer Speicherbedarf
- Historie ohne Zeitzuordnung
- Entspricht paralleler Hierarchie

Getraenk_ID	Produkt	Produktgruppe	Alte_Gruppe
1	Dornfelder	Rotwein	Rotwein
2	Müller-Thurgau	Weißwein	Weißwein
3	Portugieser Weißherbst	Rotwein	Rotwein

Getraenk_ID	Produkt	Produktgruppe	Alte_Gruppe
1	Dornfelder	Rotwein	Rotwein
2	Müller-Thurgau	Weißwein	Weißwein
3	Portugieser Weißherbst	Weißwein	Rotwein

„As is“ & „as of“-Szenario: Versionierung

- Etwas höherer Speicherbedarf
- Historie ohne Zeitzuordnung
- Künstliche Dimensionsschlüssel notwendig

Getraenk_ID	Produkt	Produktgruppe	Version	G_ID_alt	aktiv
1	Dornfelder	Rotwein	1	1	X
2	Müller-Thurgau	Weißwein	1	2	X
3	Portugieser Weißherbst	Rotwein	1	3	X

Getraenk_ID	Produkt	Produktgruppe	Version	G_ID_alt	aktiv
1	Dornfelder	Rotwein	1	1	X
2	Müller-Thurgau	Weißwein	1	2	X
3	Portugieser Weißherbst	Rotwein	1	3	
3	Portugieser Weißherbst	Weißwein	2	6	X

„As is“ & „as of“-Szenario: Zeitstempel

- Höherer Speicherbedarf
- Historie mit Zeitzuordnung
- Künstliche Dimensionsschlüssel (oder zusammengesetzte Schlüssel) notwendig
- in den Fakten Verknüpfung mit *Getraenk_ID*

G_ID	G_ID_alt	Produkt	Produktgruppe	Gültig_Von	Gültig_Bis
1	1	Dornfelder	Rotwein	01.01.2010	31.12.9999
2	2	Müller-Thurgau	Weißwein	01.01.2010	31.12.9999
3	3	Portugieser Weißherbst	Rotwein	01.01.2010	31.12.9999

G_ID	G_ID_alt	Produkt	Produktgruppe	Gültig_Von	Gültig_Bis
1	1	Dornfelder	Rotwein	01.01.2010	31.12.9999
2	2	Müller-Thurgau	Weißwein	01.01.2010	31.12.9999
3	3	Portugieser Weißherbst	Rotwein	01.01.2010	31.12.2010
3	6	Portugieser Weißherbst	Weißwein	01.01.2011	31.12.9999

„As posted“-Szenario

- Höherer Speicherbedarf
- Historie zum Transaktionszeitpunkt
- Künstliche Dimensionsschlüssel (oder zusammengesetzte Schlüssel) notwendig
- in den Fakten Verknüpfung mit *G_ID_alt*

Getraenk_ID	Produkt	Produktgruppe	Version	G_ID_alt
1	Dornfelder	Rotwein	1	1
2	Müller-Thurgau	Weißwein	1	2
3	Portugieser Weißherbst	Rotwein	1	3

Getraenk_ID	Produkt	Produktgruppe	Version	G_ID_alt
1	Dornfelder	Rotwein	1	1
2	Müller-Thurgau	Weißwein	1	2
3	Portugieser Weißherbst	Rotwein	1	3
3	Portugieser Weißherbst	Weißwein	2	6

„As is“ & „as of“ & „as posted“-Szenario: Snapshot

- Höherer Speicherbedarf in Dimensions- und Faktentabellen
- Flag für aktuelle Zuordnung
- Künstliche Ladezeitpunktdaten
- nur für “kleine“ Dimensionen empfehlenswert

Getraenk_ID	Produkt	Produktgruppe	Ladedatum	aktuell
1	Dornfelder	Rotwein	01.01.2010	0
2	Müller-Thurgau	Weißwein	01.01.2010	0
3	Portugieser Weißherbst	Rotwein	01.01.2010	0
1	Dornfelder	Rotwein	01.01.2011	1
2	Müller-Thurgau	Weißwein	01.01.2011	1
3	Portugieser Weißherbst	Weißwein	01.01.2011	1

Zusammenfassung

- Multidimensionales Datenmodell:
  - Würfel, Kennzahlen, Dimensionen
  - OLAP-Operationen
- konzeptionelle Modellierungstechniken
  - kein „Quasi“-Standard wie ER-Modell
  - spezialisierte Kontrukte für multidimensionale Konzepte
- Umsetzung des multidimensionalen Modells
  - relational: Star und Snowflake Schema
  - Mischformen
  - Vermeidung von Sematikverlusten