Data Warehousing und Data Mining

Das multidimensionale Datenmodell

Ulf Leser

Wissensmanagement in der Bioinformatik

Inhalt dieser Vorlesung

- Vom Spreadsheet zum Würfel
- Multidimensionales Datenmodell (MDDM)
- Dimensionen und Granularität
- Beispiel

Spreadsheets

- EXCEL-artige Anwendungen sind extrem weit verbreitet
  - Vorläufer der Data Warehouse Idee
  - Unverändert populär (OLAP-Plug-Ins für EXCEL)
- Beispiel: Summen von Verkaufszahlen aufgeschlüsselt nach Eigenschaften der Verkäufe

		2010												2011
		Q1			Q2			Q3			Q4			…
		Jan	Feb	Mär	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez	…
	Romane	14	45	14	14	45	24	14	45	7	64	45	14	…
Bücher	Kinder	19	13	8	19	13	8	19	13	8	19	13	8	…
	Fachliteratur	3	45	33	3	45	5	41	45	33	3	45	33	…
	Klassik	24	23	12	24	23	12	24	23	12	24	11	12	…
CDs	Hörbücher	17	5	73	17	5	73	17	5	23	17	5	73	…
	Pop	23	18	16	23	18	2	23	18	28	23	18	8	…

Spreadsheets

- EXCEL-artige Anwendungen sind extrem weit verbreitet
  - Vorläufer der Data Warehouse Idee
  - Unverändert populär (OLAP-Plug-Ins für EXCEL)
- Beispiel: Summen von Verkaufszahlen aufgeschlüsselt nach Eigenschaften der Verkäufe

Dimensionen: Zeit und Produkt

		2010												2011
		Q1			Q2			Q3			Q4			…
		Jan	Feb	Mär	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez	…
	Romane	14	45	14	14	45	24	14	45	7	64	45	14	…
Bücher	Kinder	19									19	13	8	…
	Fachliteratur	3									3	45	33	…
	Klassik	24									24	11	12	…
CDs	Hörbücher	17	5	73	17	5	73	17	5	23	17	5	73	…
	Pop	23	18	16	23	18	2	23	18	28	23	18	8	…

Detailaggregate

Hierarchische Dimensionen

Randsummen

- Wichtig sind immer auch die Randsummen
  - Jede Kombination von Dimensionen
- Schwierigkeit: Summen auf allen Hierarchiestufen
  - In EXCEL möglich, aber nicht trivial
- Vorsicht: Rand-Durchschnitte sind nicht so einfach
  - Summen schon

		2010												2011	Summe
		Q1			Q2			Q3			Q4			…	
		Jan	Feb	Mär	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez	…	
	Romane	14	45	14	14	45	24	14	45	7	64	45	14	…	345
Bücher	Kinder	19	13	8	19	13	8	19	13	8	19	13	8	…	160
	Fachliteratur	3	45	33	3	45	5	41	45	33	3	45	33	…	334
	Klassik	24	23	12	24	23	12	24	23	12	24	11	12	…	224
CDs	Hörbücher	17	5	73	17	5	73	17	5	23	17	5	73	…	330
	Pop	23	18	16	23	18	2	23	18	28	23	18	8	…	218
Summe		100	149	156	100	149	124	138	149	111	150	137	148		1611

Drei Dimensionen

- Hinzufügen einer dritten Dimension: Länder
- Erste Möglichkeit: Schachteln
  - Unklare Darstellung (dritte Dimension geht verloren)
  - Randsummen pro Land sehr unschön

		2010 …											
		Q 1									Q2…		
		Jan			Feb			Mar			Apr		
		BRD	FR	GB	BRD	FR	GB	BRD	FR	GB	BRD	FR	GB
	Romane	14	45	14	14	45	24	14	45	7	64	45	14
Bücher	Kinder	19	13	8	19	13	8	19	13	8	19	13	8
	Fachliteratur	3	45	33	3	45	5	41	45	33	3	45	33
	Klassik	24	23	12	24	23	12	24	23	12	24	11	12
CDs	Hörbücher	17	5	73	17	5	73	17	5	23	17	5	73
	Pop	23	18	16	23	18	2	23	18	28	23	18	8

Drei Dimensionen

- Hinzufügen einer dritten Dimension: Länder
- Zweite Möglichkeit: Stapeln (verschiedene Tabs)
  - Unübersichtlich, keine Gesamtschau
  - Hierarchische dritte Dimension?
  - Randsummen pro Jahr / Produkt sehr unschön

BRD

		2010											
		Q1			Q2			Q3			Q4		
		Jan	Feb	Mär	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
Bücher	Romane	14	45	14	14	45	24	14	45	7	64	45	14
	Kinder	19	13	8	19	13	8	19	13	8	19	13	8
	Fachliteratur	3	45	33	3	45	5	41	45	33	3	45	33
	Klassik	24	23	12	24	23	12	24	23	12	24	11	12
CDs	Hörbücher	17	5	73	17	5	73	17	5	23	17	5	73
	Pop	23	18	16	23	18	2	23	18	28	23	18	8

FR

		2010											
		Q1			Q2			Q3			Q4		
		Jan	Feb	Mär	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
Bücher	Romane	14	45	14	14	45	24	14	45	7	64	45	14
	Kinder	19	13	8	19	13	8	19	13	8	19	13	8
	Fachliteratur	3	45	33	3	45	5	41	45	33	3	45	33
	Klassik	24	23	12	24	23	12	24	23	12	24	11	12
CDs	Hörbücher	17	5	73	17	5	73	17	5	23	17	5	73
	Pop	23	18	16	23	18	2	23	18	28	23	18	8

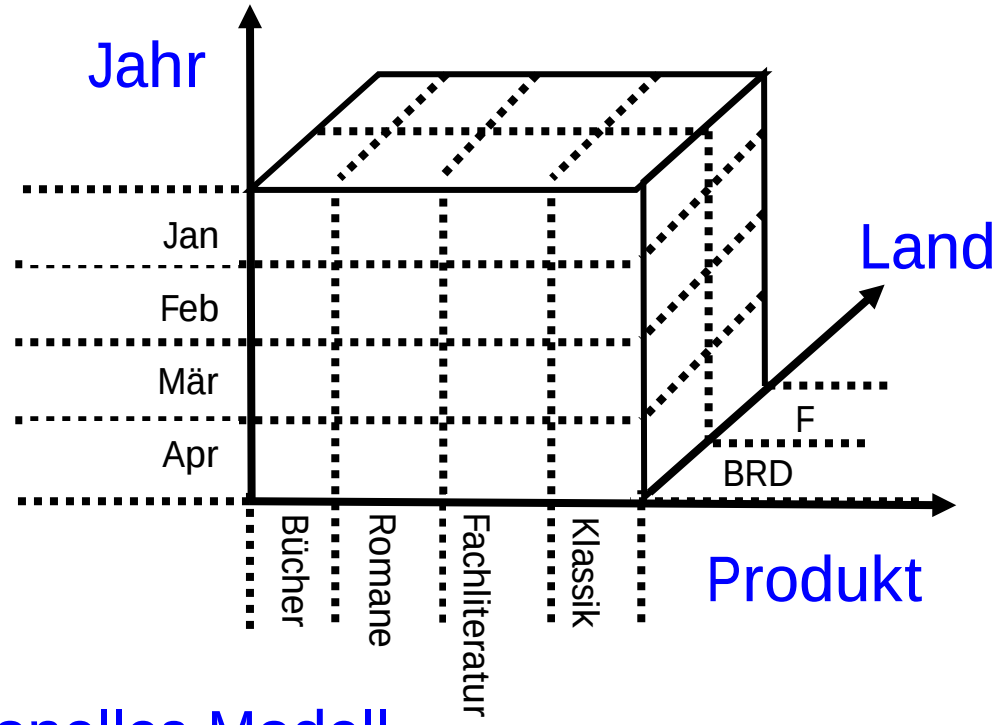
GB

		2010											
		Q1			Q2			Q3			Q4		
		Jan	Feb	Mär	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
Bücher	Romane	14	45	14	14	45	24	14	45	7	64	45	14
	Kinder	19	13	8	19	13	8	19	13	8	19	13	8
	Fachliteratur	3	45	33	3	45	5	41	45	33	3	45	33
	Klassik	24	23	12	24	23	12	24	23	12	24	11	12
CDs	Hörbücher	17	5	73	17	5	73	17	5	23	17	5	73
	Pop	23	18	16	23	18	2	23	18	28	23	18	8

Nachteile von Spreadsheets

- Bei mehr als zwei Dimensionen unübersichtlich
  - Mehr als drei Dimensionen kaum handhabbar
- Schwierigkeiten mit hierarchischen Dimensionen
- Schwierigkeiten mit spezielleren Anforderungen
  - Gleitenden Durchschnitt pro Ort über je 3 Jahre?
  - Immer Summe und Durchschnitt anzeigen?
- Starke Beschränkung in möglicher Datenmenge
  - Wir haben nur Aggregate gezeigt – Originaldaten?
- Verwirrende Informationsvielfalt
  - Keine Selektionen, Views, ...

Lösung



- Konzeptionelles Modell
  - Kann man nicht ausdrucken – stapeln oder schachteln
  - Wird in einem UI in Ausschnitten angezeigt
  - UI / DM- Operationen arbeiten auf dem Cube und erzeugen eine Ansicht

Inhalt dieser Vorlesung

- Vom Spreadsheet zum Würfel
- Multidimensionales Datenmodell (MDDM)
- Dimensionen und Granularität
- Beispiel

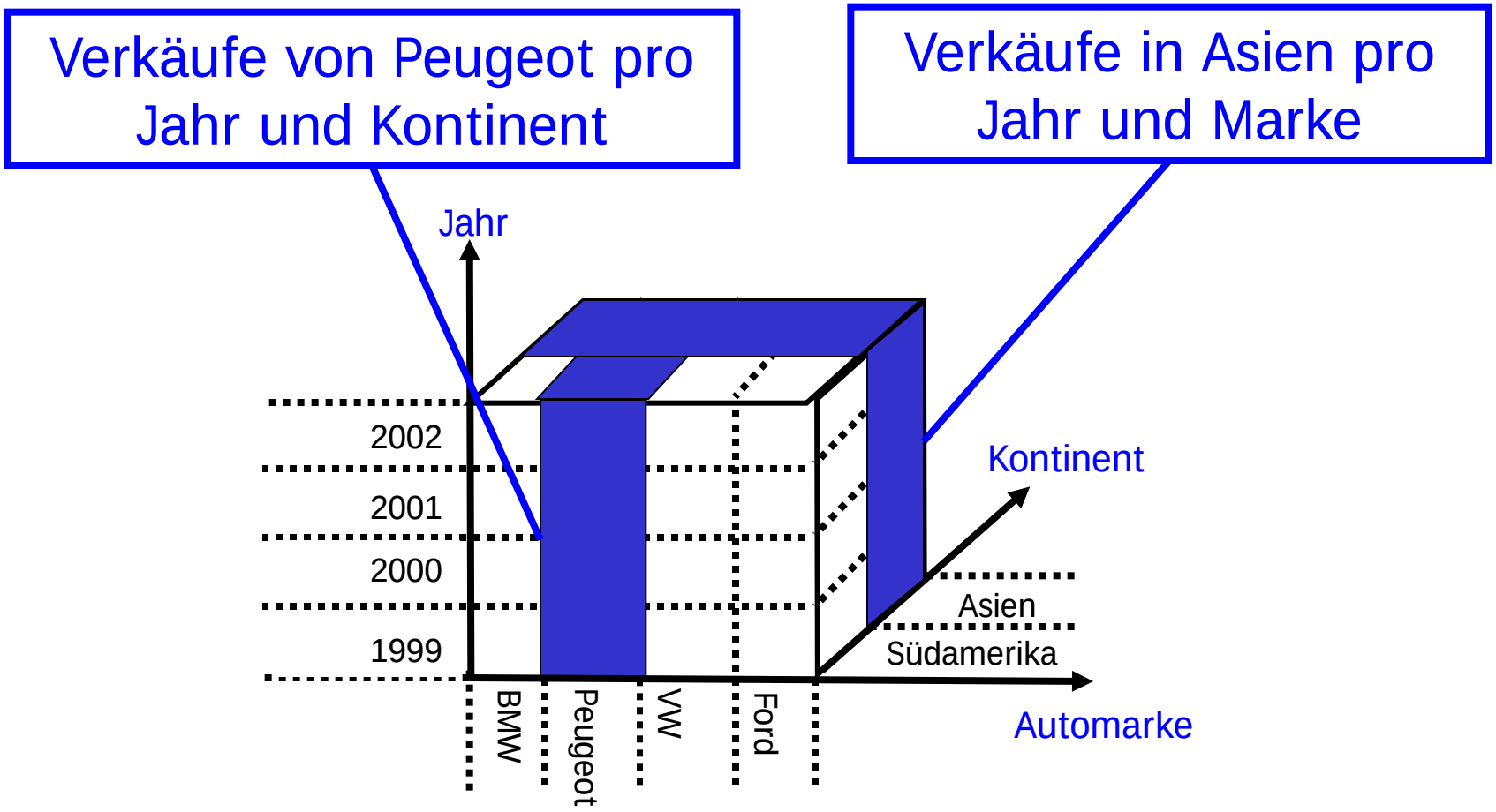
MDDM Grundidee

- Wichtigste Elemente
  - Fakten (Measures) – Gemessene Werte
  - Dimensionen – Beschreibung der Fakten in Raum, Zeit, …
  - Klassifikationshierarchien – Strukturierte Dimensionen
- Metapher: Würfel (Cube) bzw. Hypercube
  - Fakten: Punkte im multidimensionalen Raum
  - Dimensionen: Achsenbeschriftung / Koordinaten
  - Hierarchien: Koordinaten in unterschiedlichem Detailgrad
- Analyse durch Operationen auf dem Cube
  - Auswahl von Subwürfeln (Flächen, Punkten, …)
  - Hierarchiestufe vergröbern/verfeinern
    - Mit Aggregation der Daten
  - Nächste Stunde

Beispiel

- Verkäufe von Autos pro Marke, Kontinent und Jahr gemessen in Euro
  - Fakten
    - Verkäufe in Euro
  - Dimensionen
    - Automarke
    - Kontinent
    - Jahr

Beispiel: Auswahl (Slicing)



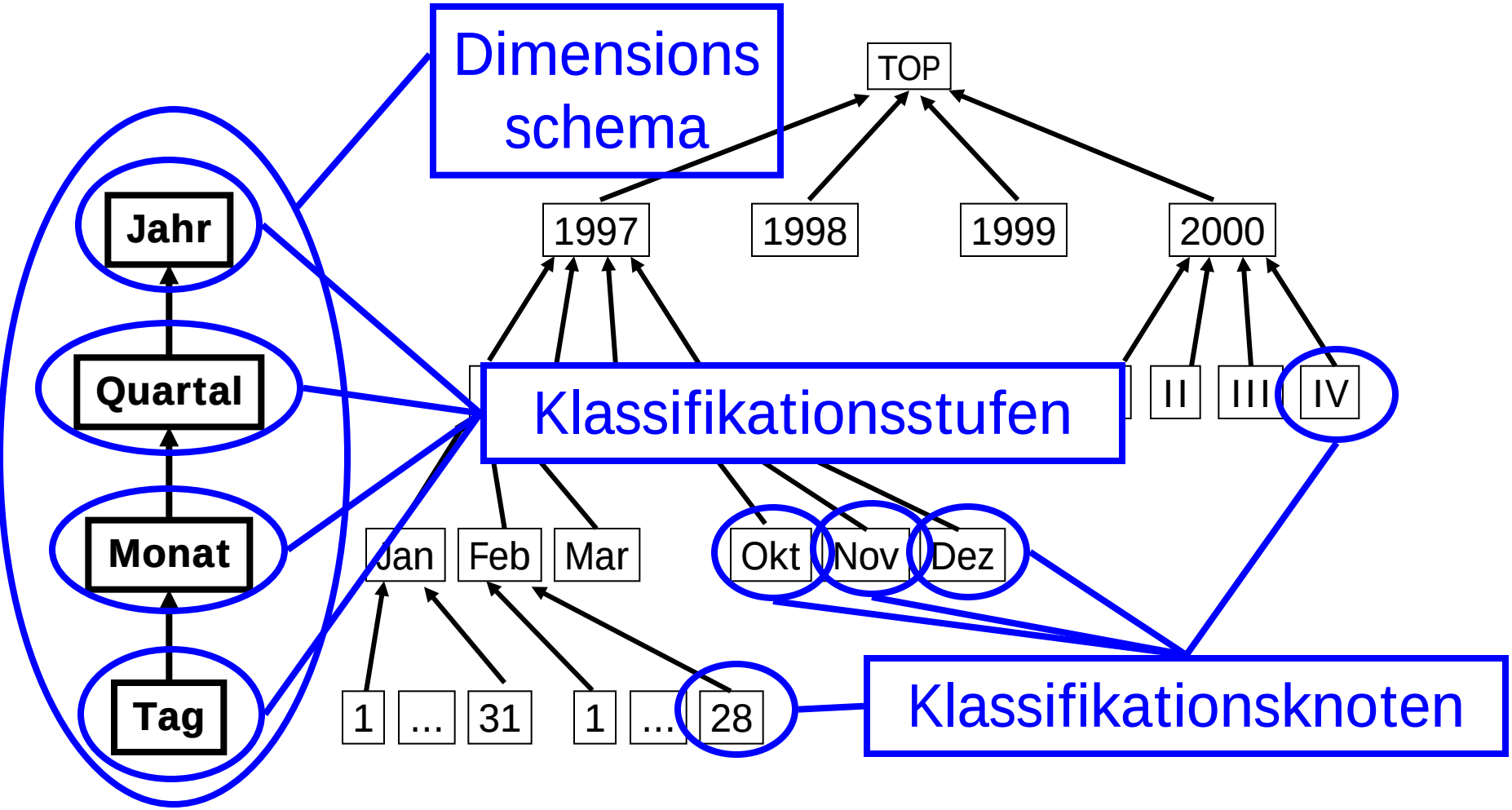
Fakten (Measures)

- Fakten haben Koordinaten in allen Dimensionen
  - Partielle Koordinaten: DQ Problem, Imputation
- Fakten haben Werte
  - Können auch mehrere sein, z.B. Uhrzeit, Verkaufswert
- Wichtigste Unterscheidung
  - Event facts: Fakten entsprechen Ereignissen der realen Welt
    - Messungen, Verkäufe, Wählerstimmen (am Wahltag), ...
    - Können über die Zeit aggregiert werden
      - Summe, Mittelwert, Quantil, ...
  - Snapshot facts: Fakten entsprechen Zuständen der realen Welt
    - Lagerbestände
    - Können nicht über die Zeit summiert werden
      - Durchschnitt macht aber Sinn
    - Aber über alle anderen Dimensionen

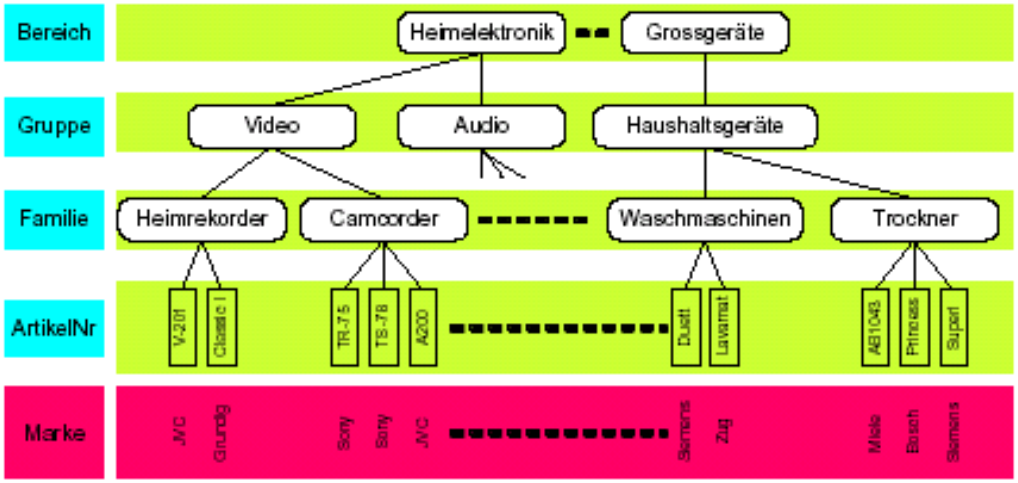
Dimensionen

- Beschreiben relevante Eigenschaften von Fakten
- Müssen orthogonal (funktional unabhängig) sein
  - Abhängigkeiten bereiten viele Probleme – später
- Haben meist ein hierarchisches Schema
  - Zeit: Tag, Woche, Jahr, ...
    - Achtung: Es sind individuelle Tage, Monate, ... gemeint. Als nicht „1 = der erste jedes Monats“, sondern „ID1 = 1.1.2006“ etc.
  - Region: Landkreis, Land, Staat, ...
  - Produkt: Produktgruppe, Produktklasse, Produktfamilie, ...
- Haben eine erlaubte Wertemenge auf jeder Ebene
  - (1/1/99, 2/1/99, ..., 31/1/99,...31/12/07), ...
  - (Berlin, NRW, Department-1, ...), (BRD, F, ...)
- Das ergibt ein Schema

Dimension


Dimensions schema
TOP
1997
1998
1999
2000
Jahr
Quartal
Monat
Tag
Klassifikationsstufen
II
III
IV
Jan
Feb
Mar
Okt
Nov
Dez
1
...
31
1
...
28
Klassifikationsknoten

Produkthierarchie

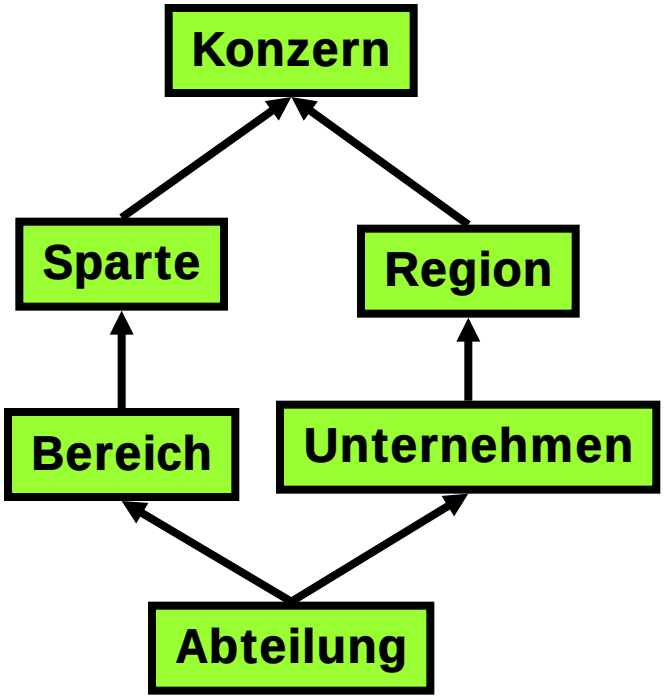


Aus: Geppert, ETZ Zürich, Vorlesung „Data Warehouse“

- Elemente einer Stufe können geordnet sein
  - Geordnet: Zeit
  - Ungeordnet: Produkte

Formale Definition MDDM

- Ziele
  - Operationen im MDDM exakt definieren
  - Aus dem Modell erkennen, welche Aggregate semantisch sinnvoll sind und welche nicht
    - Tools bieten dann nur die sinnvollen Operationen an
  - Systeme können Informationen zur Anfrageoptimierung benutzen
  - Grafische Spezifikation und Visualisierung



Klassifikationsschema (einer Dimension)

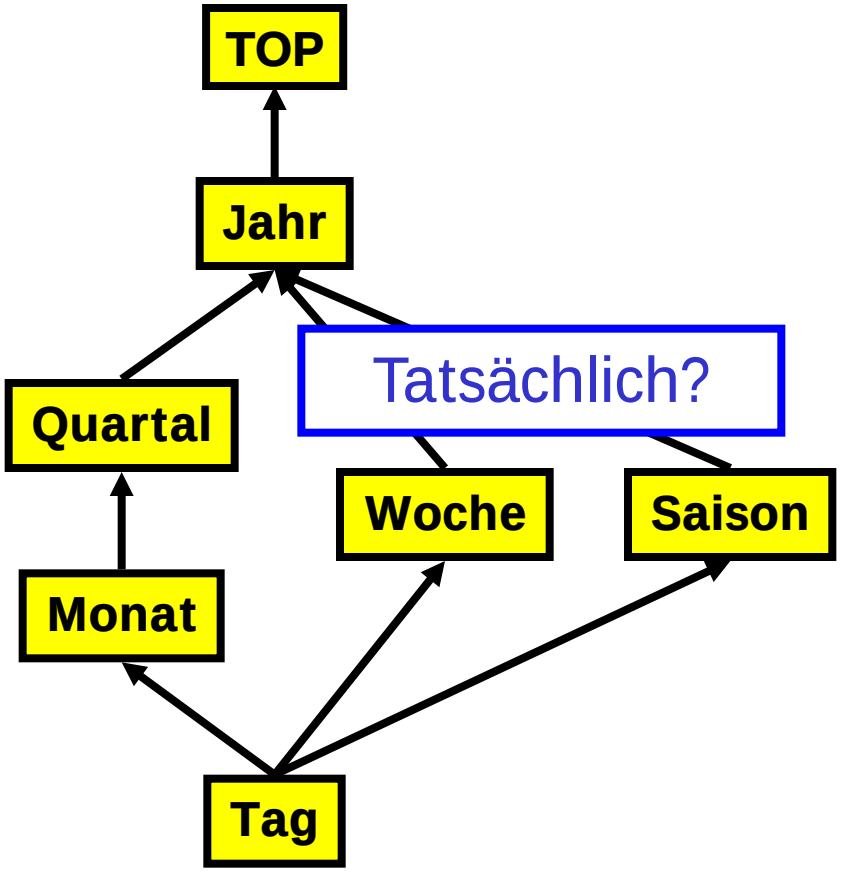
- Definition
  Ein Klassifikationsschema K ist ein Quadrupel
  $(K_s, \rightarrow, K_k,$ stufe) mit
  - *K_s ist die Menge von Klassifikationsstufen $\{k_0, \ldots k_n\}$*
  - *„$\rightarrow$“ ist eine transitive, azyklische und antisymmetrische Relation zwischen zwei K-stufen mit größtem Element top(K_s)*
    - *D.h.: $\forall k \in K_s: k \rightarrow_s top(K_s)$*
  - *K_k ist die Menge von Klassifikationsknoten $\{n_0, \ldots n_m\}$*
  - *stufe ordnet jedem K-Knoten n genau eine K-Stufe k zu*
- Bemerkung
  - Eine Klassifikationsstufe hat mehrere Klassifikationsknoten, aber jeder Klassifikationsknoten ist genau einer Klassifikationsstufe zugeordnet
  - Man definiert dazu Funktion knoten(i) = $\{n \mid n \in K_k \wedge \text{stufe}(n)=i\}$

Erläuterung

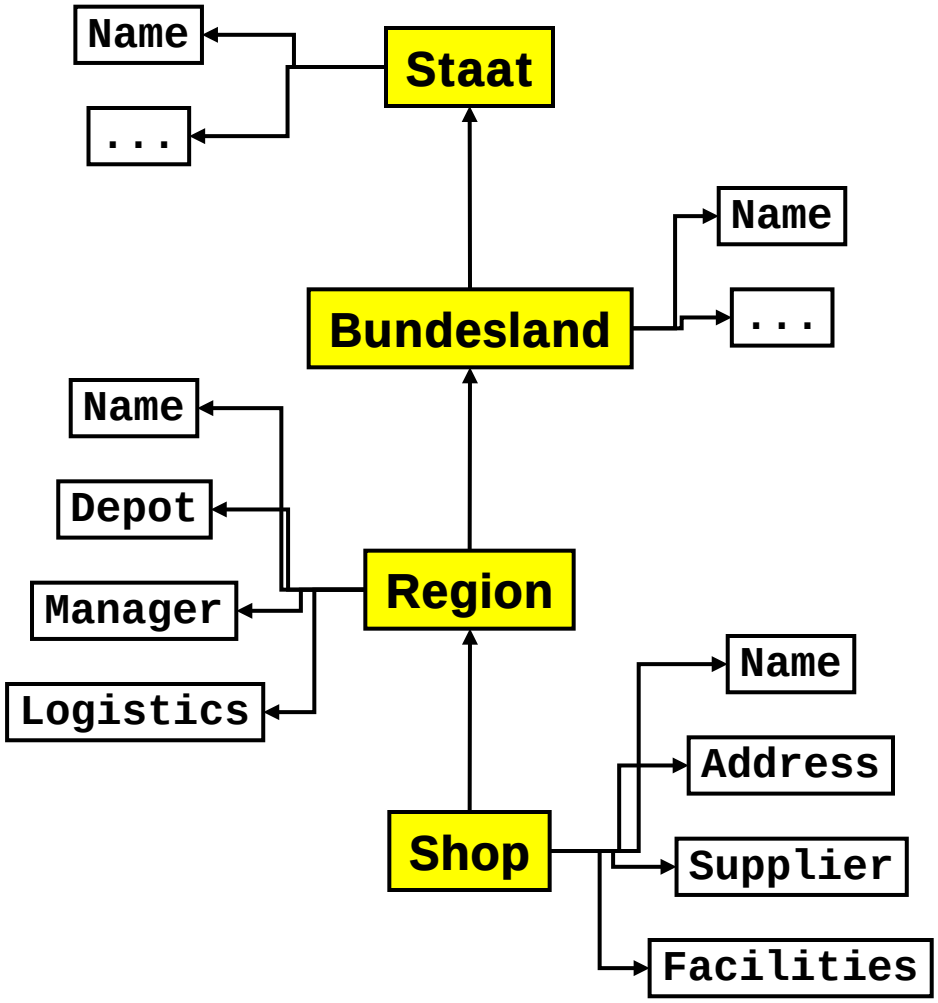
- Klassifikationsstufen: Schemaelemente der Dimension
- Klassifikationsknoten: Instanzen der Schemaelemente
- Das größte Element (top) ist artifiziell
  - Steht für „alles“
- Interpretation von „→“
  - Funktionale Abhängigkeit & Aggregierbarkeit
  - Tag bestimmt Monat bestimmt Jahr bestimmt TOP
    - 21.12.2003 → 12.2003 → 2003 → TOP
  - Produkt → Produktfamilie → Produktgruppe → TOP
    - “Asus M2400N“ → Notebooks → Büroelektronik → TOP

Beispiel

- Ordnung
  - Tag → Monat
  - Monat → Quartal
  - Quartal → Jahr
  - Tag → Woche
  - Woche → Jahr
- Partielle Ordnung
  - Quartal ? Woche
  - Monat ? Woche
- Transitivität
  - Tag → Jahr
  - Alle → TOP



Knotenattribute



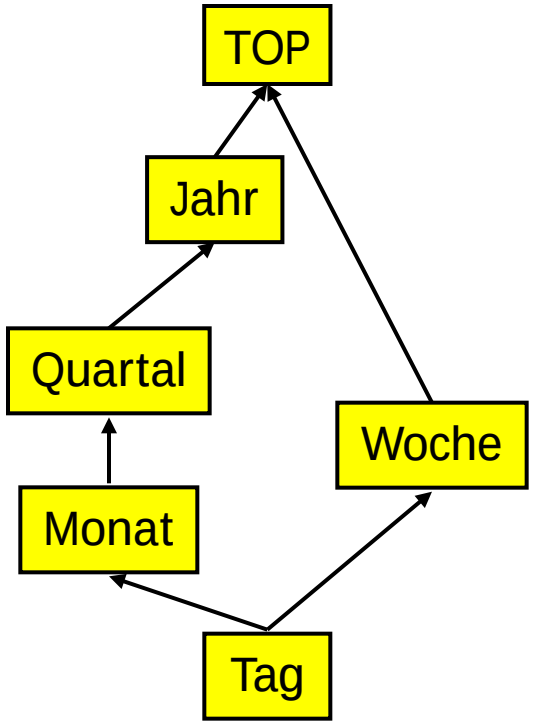
- Jede Klassifikationsstufe hat eine Menge von Attributen, die Knotenattribute
  - Teil des Klassifikationsschemas

Klassifikationspfade

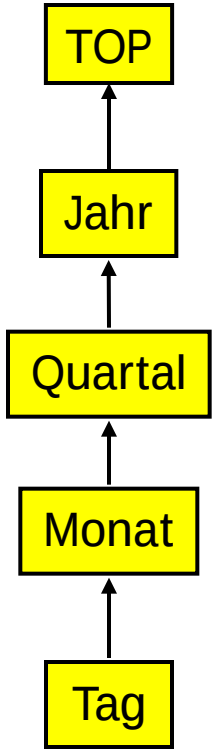
- Definition
  Ein Klassifikationspfad P in einem Klassifikationsschema K mit Klassifikationsstufen K_s ist eine Menge $\{p_0, \ldots, p_m\} \subseteq K_s$ mit
  - $p_m = top(K_s)$
  - $\forall p_i , 1 \leq i \leq m: p_{i-1} \rightarrow p_i$ und $\nexists q: p_{i-1} \rightarrow q \rightarrow p_i$
  - *Die Länge des Pfades P ist $|P|=m+1$*
  - *Der Klassifikationslevel von p_i in P ist i*
- Bedeutung
  - P: Zusammenhängende und vollständig geordnete Teilmenge von K_s
  - Aggregate werden wir entlang von Pfaden definieren
    - Und damit entlang funktionaler Abhängigkeiten

Beispielpfade

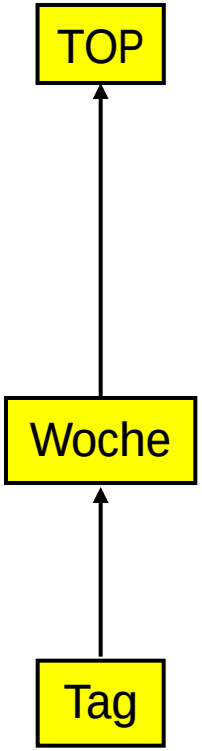
Klassifikationsschema



Pfad 1

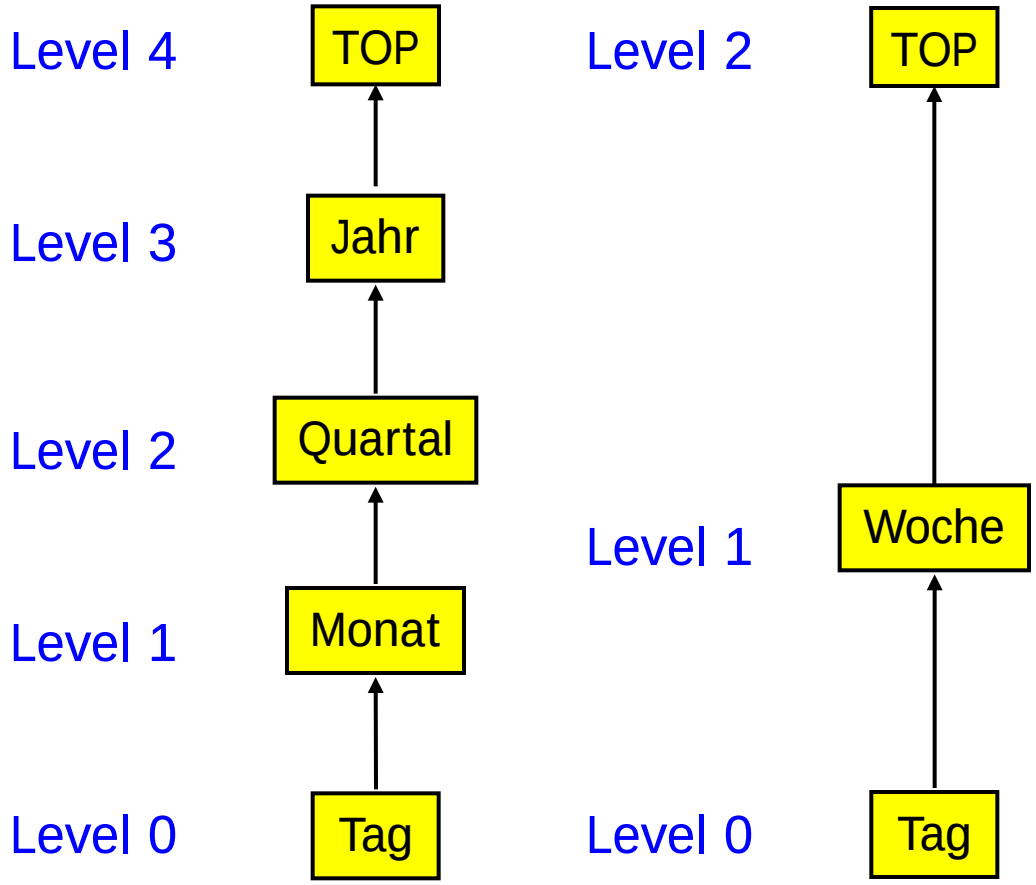


Pfad 2



Klassifikationsstufen und -level

- Der Klassifikations-level einer Stufe ist nur eindeutig in einem Pfad
- Der Level des größten Elements TOP ist daher nicht konstant
  - Wir zählen von unten



Klassifikationshierarchie

- Definition
  - *Die Klassifikationshierarchie H zu einem* Klassifikationsschema *K mit Pfad P ist der Baum mit Knoten N und Kanten E wie folgt:*

$$N = \bigcup_{p_i \in P} knoten(p_i)$$

$$E = \left\{ (n,m) \;\middle|\; \begin{array}{l} n,m \in N \wedge stufe(n) \rightarrow stufe(m) \wedge \\ \qquad \exists j : n \in knoten(p_j) \wedge m \in knoten(p_{j+1}) \end{array} \right\}$$

- Bemerkungen
  - Jede Klassifikationshierarchie ist balanciert: Alle Pfade Wurzel-Blatt haben die Länge |P|

Inhalt dieser Vorlesung

- Vom Spreadsheet zum Würfel
- Multidimensionales Datenmodell (MDDM)
- Dimensionen und Granularität
- Beispiel

Dimension

- Definition
  Eine Dimension $D=(K, \{P_1,\ldots,P_j\})$ besteht aus
  - *Einem Klassifikationsschema K*
  - *Einer Menge von Pfaden P_i in K*
- Bemerkungen
  - D muss nicht alle Pfade enthalten, die es in K gibt
  - Theoretisch müssen nicht alle Klassifikationsstufen von K in (mindestens) einem Pfad enthalten sein
    - Aber man wird seine Pfade so wählen, dass dies doch gilt
- Schreibweise
  - **`D.k`** bezeichnet die Klassifikationsstufe **`k`** aus **`D`**
  - Jedes **`D.k`** kann in mehreren Pfaden vorkommen

Granularität

- Definition
  - *Ein Dimensionsschema U ist eine Menge von Dimensionen $D_1,\ldots, D_n$ für die gilt, dass es keine funktionalen Abhängigkeiten zwischen den Klassifikationsstufen verschiedener Dimensionen gibt.*
  - *Eine Granularität G über einem Dimensionsschema U ist eine Menge $\{D_1.k_{i1},\ldots ,D_n.k_{in}\}$ mit $\forall j$: k_{ij} ist Klassifikationsstufe in D_i*
- Bemerkungen
  - Funktionale Abhängigkeiten sind semantisches Wissen; die kann das DWH nicht erkennen
  - Beispiel: Dimensionen Zeit und „Fiskalisches Jahr“ können nicht gleichzeitig in einem Dimensionsschema verwendet werden
    - Sonst kann man sinnlose Granularitäten / Gruppierungen bilden

Erläuterung

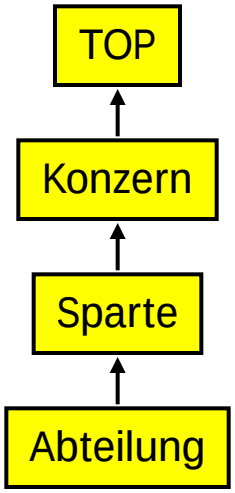
- Mit einer Granularität legt man fest, in welcher Detailstufe Daten betrachtet werden
  - Festlegung für jede Dimension
  - Würfel in einer bestimmten Auflösung
- OLAP-Operationen operieren auf Granularitäten und erzeugen (andere) Granularitäten
  - Navigation entlang von Pfaden
  - Selektion von Knoten
  - Herausschneiden von Dimensionen

Halbordnung auf Granularitäten

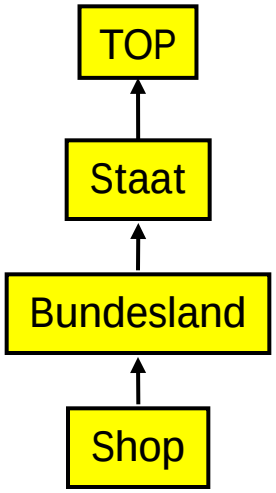
- Durch die Anordnung der Stufen in Pfaden sind Granularitäten halb-geordnet
- Definition
  Auf der Menge aller Granularitäten eines Dimensions-schemas U ist eine Halbordnung „≤“ wie folgt definiert
  - *Sei* $G_1=\{D_1.k_{i1}^1,\ldots,D_n.k_{in}^1\}$ *und* $G_2=\{D_1.k_{i1}^2,\ldots,D_n.k_{in}^2\}$
  - *Es gilt:* $G_1 \leq G_2$ *genau dann wenn* $\forall j: D_i.k_{ij}^1 \rightarrow D_i.k_{ij}^2$
- Bemerkung
  - $D_i.k_i^1 \rightarrow D_i.k_i^2$ ist auch erfüllt wenn $D_i.k_i^1 = D_i.k_i^2$
  - Die Halbordnung repräsentiert erlaubte Transformation von Granularitäten
  - Anfrageoptimierung: Wiederverwendung von Aggregaten

Beispiel

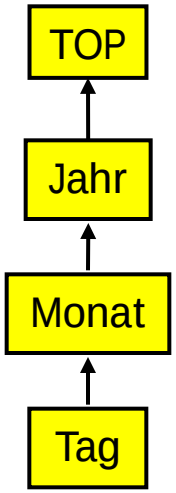
Bereich



Region



Zeit



(B.Sparte, R.Shop, Z.Tag)
≤ (B.Sparte, R.Shop, Z.Monat)
≤ (B.Sparte, R.TOP, Z.Monat)
≤ (B.TOP, R.TOP, Z.TOP)

(B.Sparte, R.Staat, Z.Tag)
? (B.Konzern, R.Shop, Z.Tag)

Würfelschema und Würfel

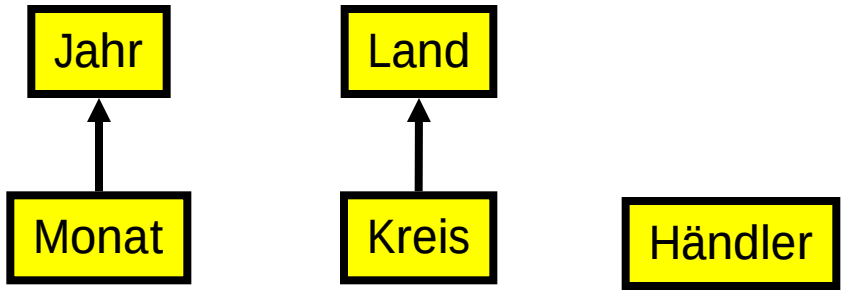
- Definition
  Ein Würfelschema WS ist ein Tupel (G,F) mit
  - *Einer Granularität G*
  - *Einer Menge F von Fakten mit |F|=m*
- *Ein Würfel W ist eine vollständige Instanz eines Würfelschema (G,F)*

$$W = dom(G) \times dom(F)$$
$$= knoten(D_1.k_1) \times ... \times knoten(D_n.k_n) \times dom(F_1) \times ... \times dom(F_m)$$

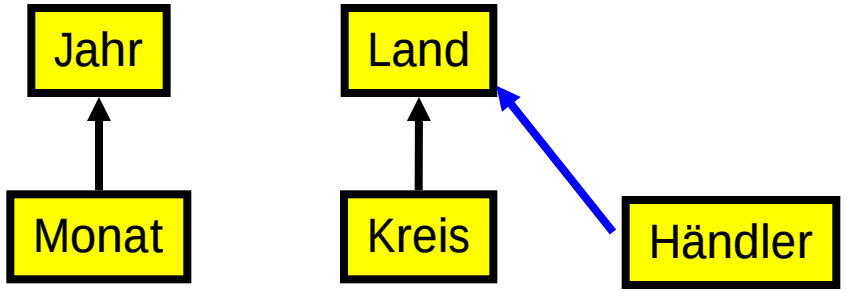
- Bemerkung
  - Die Werte dom(G) geben die Koordinaten der Werte dom(F) an
  - Würfelschema zu Würfel ist wie Relationenschema zu Relation

Kein Würfelschema

- Autoverkäufe pro Zeit (Monat, Jahr), Händler und Region (Kreis, Land)
- Drei Dimensionen
  - Monat → Jahr
  - Händler
  - Kreis → Land

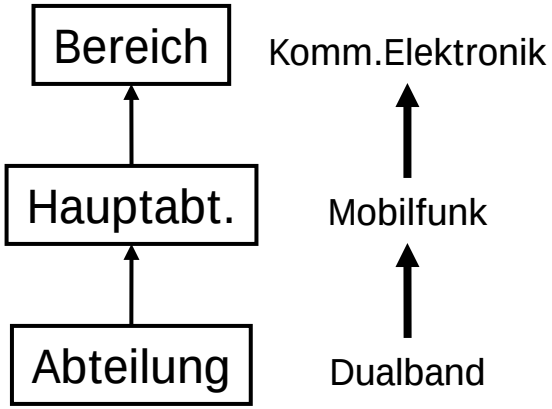


- Aber: EU Recht !
  - Händler → Land
  - Damit könnten wir sinnlose Granularitäten bauen
  - Alternative: Dimensionen anders definieren

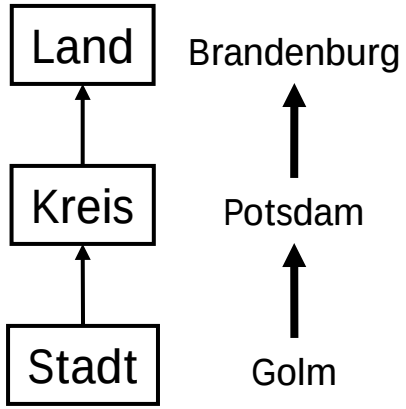


Semantik von Kanten

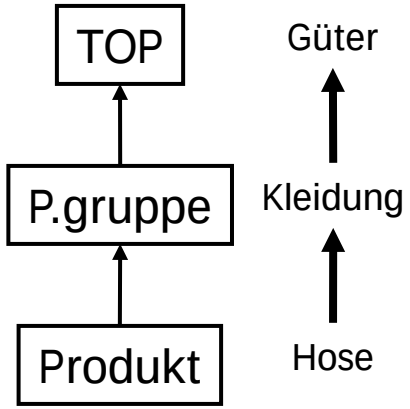
- Die Hierarchie von Klassifikationsstufen wird durch funktionale Abhängigkeiten bestimmt
- Das beinhaltet keine Aussage zur Semantik von Kanten



gehört_zur_ organisationseinheit



geographisches PART_OF



IS-A

Inhalt dieser Vorlesung

- Vom Spreadsheet zum Würfel
- Multidimensionales Datenmodell (MDDM)
- Dimensionen und Granularität
- Beispiel

Ein längeres Beispiel

- Wir bauen ein DWH zur Verwaltung von Lagerbeständen
- Wir haben viele Lagerhäuser (international verteilt)
- „Messung“ ist der Zugang oder Abgang von Produkten
- Jede Messung erzeugt zwei Fakten
  - Bestand und Delta von Artikeln
  - Nur über das eine darf man in der Zeit aggregieren
- Klassifikationsschema K
  - Zeit: Monat, Quartal, Woche, Jahr
  - Ort: Region, Land
  - Produkt: Artikel, Artikelgruppe, Bereich

Klassifikationsknoten

- Jahr: 1997, 1998, 1999, ...
- Quartal: I, II, III, IV (pro Jahr)
- Woche: 1-52 (pro Jahr)
- Monate: 1-3 (pro Quartal I), 4-6 (pro Quartal II), ...
- Land: Deutschland, Frankreich, Großbritannien, ...
- Region: Bayern, Berlin, ..., Departament-1, Departament-2, ...
- Bereich: Kleidung, Nahrung, Elektronik, ...
- Artikelgruppe: Oberbekleidung, Unterbekleidung, Kindernahrung, Kleingeräte, TV/Video, ...
- Artikel: ...

- Alle existierenden Ausprägungen der K-stufen

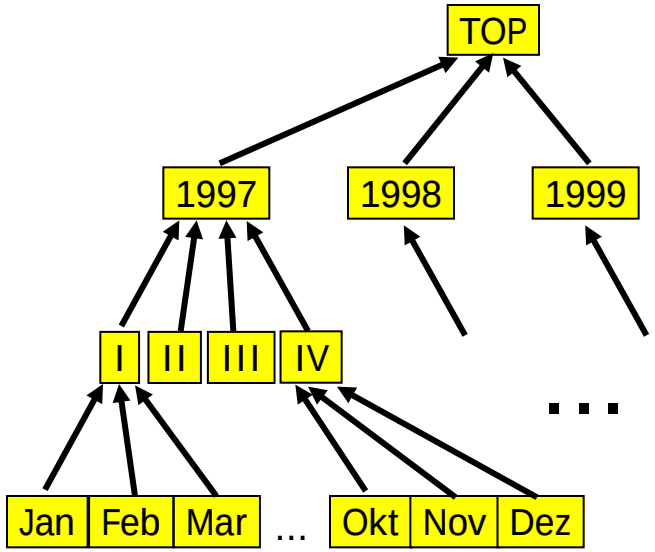
Pfade

- P_1: TOP ← Jahr ← Quartal ← Monat
- P_2: TOP ← Woche
- P_3: TOP ← Land ← Region
- P_4: TOP ← Bereich ← Artikelgruppe ← Artikel

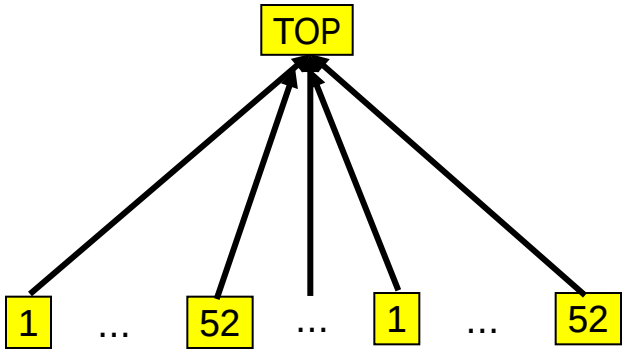
- Entlang der Pfade sind Verdichtungen im Modell sinnvoll

Klassifikationshierarchien

- Klassifikationshierarchie zu P_1
  - P_1: TOP ← Jahr ← Quartal ← Monat



- Klassifikationshierarchie zu P_2
  - P_2: TOP ← Woche



Dimensionen

- Dimension ZEIT
  - ({Monat, Quartal, Woche, Jahr}, {P_1, P_2})
- Dimension ORT
  - ({Region, Land}, {P_3})
- Dimension PRODUKT
  - ({Artikel, Artikelgruppe, Bereich}, {P_4})
- Dimensionsschema: {ZEIT, ORT, PRODUKT}

- Dimensionen enthalten mehrere Pfade

Granularität, Würfel

- Mögliche Granularitäten
  - G_1 = (Zeit.Woche, Ort.Land, Produkt.Artikel)
  - G_2 = (Zeit.Jahr, Ort.Gebiet, Produkt.TOP)
  - Halbordnung:
    - (Zeit.Tag, Ort.Gebiet, Produkt.Artikel)
    - ≤ (Zeit.Jahr, Ort.Gebiet, Produkt.Bereich)
    - ≤ (Zeit.Jahr, Ort.TOP, Produkt.Bereich)
    - ≤ (Zeit.TOP, Ort.TOP, Produkt.TOP)
- Würfelschema
  - Granularität plus Menge von Fakten (F_1=Bestand, F_2=Delta)
- Würfel: Instanz des Würfelschemas
  - Konkrete Bewegungen von Produkten aggregiert auf einer bestimmten Detailstufe pro Dimension

Weiterführende Literatur

- Vassiliadis, P. (1998). "Modeling Multidimensional Databases, Cubes, and Cube Operations". 10th International Conference on Scientific and Statistical Database Management, Capri, Italy.
- Rizzi, S., Abello, A., Lechtenbörger, J. and Trujillo, J. (2006). "Research in Data Warehousing Modeling and Design: Dead or Alive?" DOLAP, Arlington, USA.
- Vetterli, T., Vaduva, A. and Staudt, M. (2000). "Metadata Standards for Data Warehousing: Open Information Model versus Common Warehouse Model." *SIGMOD Record 29(3): 68-75.*

Selbsttest

- Was sind Klassifikationsknoten, Klassifikationsstufen, Klassifikationshierarchie?
- Warum ist eine Klassifikationshierarchie in unserem formalen Modell ein balancierter Baum?
- Wie haben wir eine Halbordnung auf Granularitäten definiert? Wozu ist die gut?
- Nennen Sie ein paar Unterschiede in der Philosophie und Verwendung des relationalen Datenmodells versus des MDDM
- Wie kann man mehrdimensionale Daten (d>3) geeignet ausdrucken?