



**Hasso
Plattner
Institut**

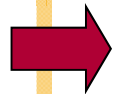
IT Systems Engineering | Universität Potsdam

Datenbanksysteme I
Data Warehouses

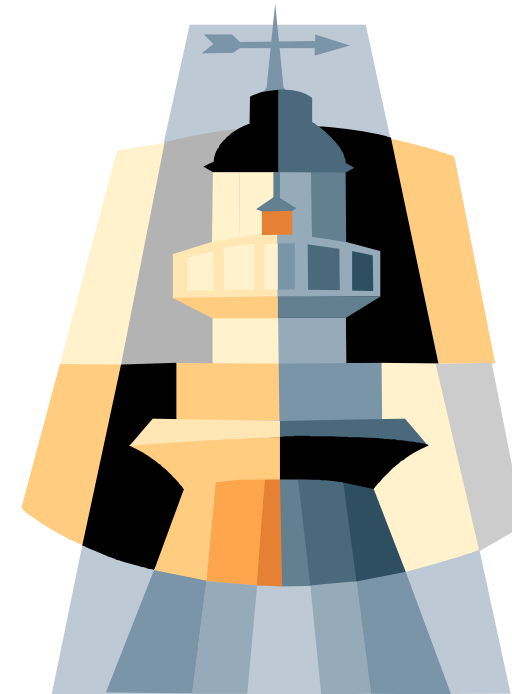
5.7.2010

Felix Naumann

2



- Einsatzgebiete
- OLAP versus OLTP
- Multidimensionale Modellierung
- OLAP Operationen
- Relationale Implementierung
- Neue Architekturen



Folien zu DWH teilweise von: Ulf Leser (HU Berlin)

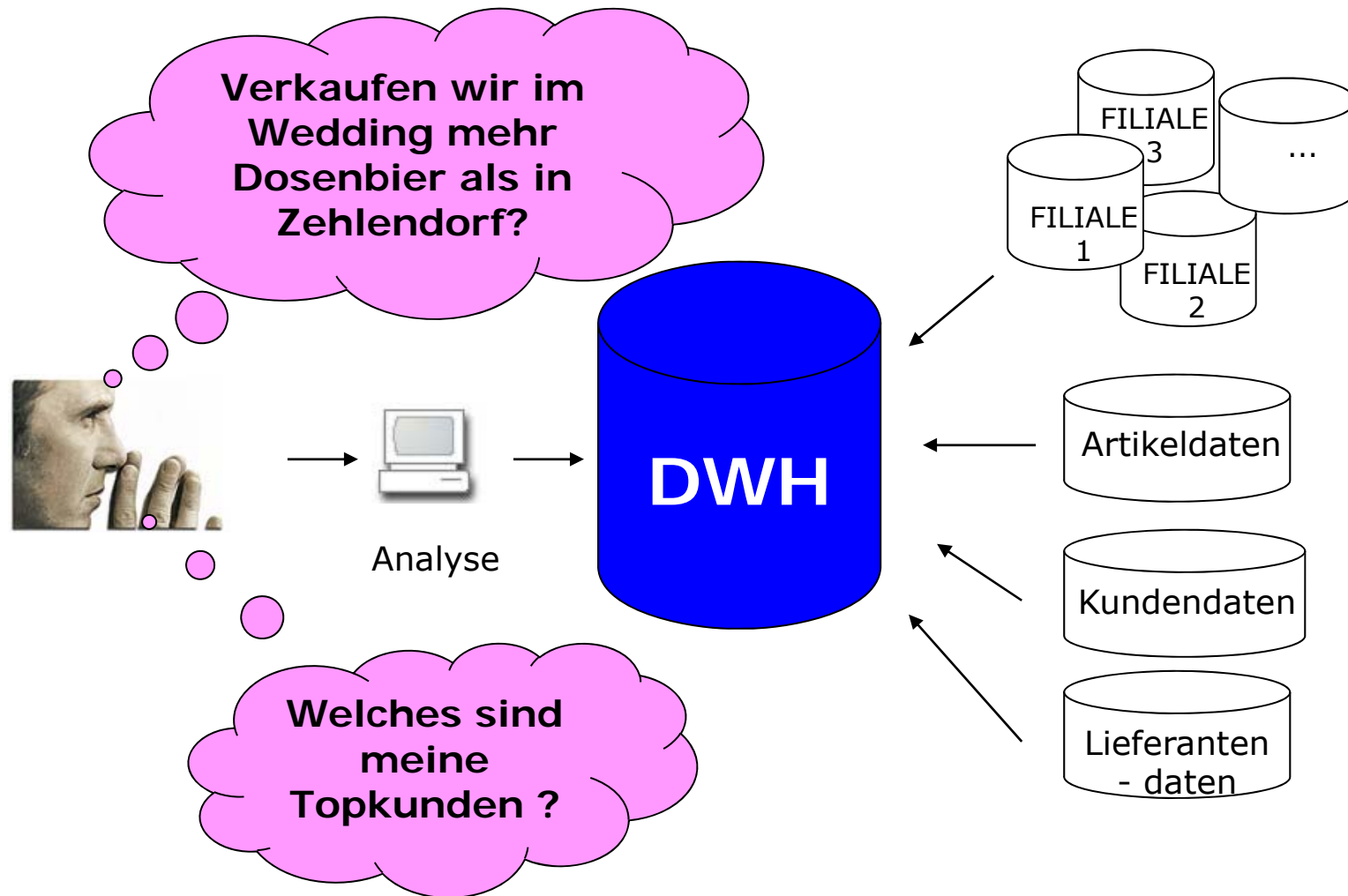
Beispielszenario

3

- Ein beliebiges Handelshaus: Spar, Kaufland, ...
- Physikalische Datenverteilung
 - Viele Niederlassungen (bis zu mehrere tausend)
 - Noch mehr Registerkassen
- Aber: Zentrale Planung, Beschaffung, Verteilung
 - Was wird wo und wie oft verkauft?
 - Was muss wann wohin geliefert werden?
 - ◇ Bedenke: Verderbliche Waren
- ... nur möglich, wenn
 - Zentrale Übersicht über Umsätze
 - Integration mit Lieferanten / Produktdaten

Handels - DWH

4



DWH Datenquellen

5

- Lieferantendatenbanken
 - Produktinformationen: Packungsgrößen, Farben, ...
 - Lieferbedingungen, Rabatte, Lieferzeiten, ...
- Personaldatenbank
 - Zuordnung Kassenbuchung auf Mitarbeiter
 - Stundenabrechnung, Prämien
- Kundendatenbank
 - Kundenklassen: Premium, normal, soziale Brennpunkte, ...
 - Persönliche Vorlieben & Historie
 - ◇ Kundenkarten (Safeway, ...)
- Weitere Vertriebswege
 - Internet, Katalogbestellung, Verkaufclubs, ...

Definition DWH

6

A DWH is a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management's decisions.

[Inman96]

- Subject-oriented: Verkäufe, Personen, Produkte, etc.
- Integrated: Erstellt aus vielen Quellen
- Non-Volatile: Hält Daten unverändert über die Zeit
- Time-Variant: Vergleich von Daten über die Zeit

Erfolg von DWH

7

- Top-Thema seit Mitte der 90er Jahre
 - Industrie schneller als Forschung
- Voraussetzungen
 - Extreme Verbilligung von Plattenspeicherplatz
 - Relationale Modellierung: Anwendungsneutral
 - Graphische Benutzeroberflächen und Terminals
 - IT in allen Unternehmensbereichen (SAP R/3)
 - Vernetzung und DB Standardisierung (SQL)
- Aber
 - Vision der vollständigen Integration scheitert (immer wieder aufs neue)
 - ◇ ETL
 - Soziale versus technische Aspekte

Datenbank größen

8

2005 TopTen Award Winners

Winter Corporation recognizes these organizations and their vendors for their achievements in the 2005 TopTen Program.

[List of all the winners](#) [Frequently Asked Questions](#)

Pick a TopTen Award Category:

Metric: Norm. Data Volume **Platform:** All Environments **Usage:** DW

Norm. Data Volume, All Environments, DW *

Company/Organization	Norm. Data Volume (GB)	DBMS	Platform	Architecture	DBMS Vendor	System Vendor	Storage Vendor
AT&T	330,644	Daytona	UNIX	Federated/SMP	AT&T	HP	HP
AT&T	93,468	Daytona	UNIX	Federated/SMP	AT&T	Sun	Sun
Amazon.com	28,184	Oracle RAC	Linux	Centralized/Cluster	Oracle	HP	HP
Nielsen Media Research	17,969	Sybase IQ	UNIX	Centralized/SMP	Sybase	Sun	EMC
Yahoo!	17,014	Oracle	UNIX	Centralized/SMP	Oracle	Fujitsu Siemens	EMC
Amazon.com	14,849	Oracle RAC	Linux	Centralized/Cluster	Oracle	HP	HP
UBS AG	14,177	Oracle	UNIX	Centralized/SMP	Oracle	Sun	EMC
China Telecom Corporation Co.,Ltd. GuangZhou Research Institute	13,241	Sybase IQ	UNIX	Centralized/SMP	Sybase	Sun	Sun
USDA	11,876	SQL Server	Windows	Centralized/SMP	Microsoft	HP	EMC
Reliance Infocomm Ltd	11,500	Oracle	UNIX	Centralized/SMP	Oracle	Sun	EMC
Cellcom	10,345	Oracle RAC	UNIX	Centralized/Cluster	Oracle	HP	EMC

Normalized Data Volume estimates of the total volume of data managed by the DBMS in GB.

- http://www.wintercorp.com/VLDB/2005_TopTen_Survey/TopTenWinners_2005.asp

Date
-größe

9

2005 TopTen Award Winners

Winter Corporation recognizes these organizations and their vendors for their achievements in the 2005 TopTen Program.

[List of all the winners](#)

[Frequently Asked Questions](#)

Pick a TopTen Award Category:

Metric: Peak Workload
Platform: All Environments
Usage: DW

Number of Rows, All Environments, DW *

Company/Organization	Number of Rows (Millions)	DBMS	Platform	Architecture	DBMS Vendor	System Vendor	Storage Vendor
Sprint	2,847,553	NonStop SQL	NonStop OS	Centralized/Cluster	HP	HP	HP
AT&T	1,882,638	Daytona	UNIX	Federated/SMP	AT&T	HP	HP
AT&T	533,723	Daytona	UNIX	Federated/SMP	AT&T	Sun	Sun
Nielsen Media Research	502,407	Sybase IQ	UNIX	Centralized/SMP	Sybase	Sun	EMC
Yahoo!	385,318	Oracle	UNIX	Centralized/SMP	Oracle	Fujitsu Siemens	EMC
Turkcell	181,083	Oracle	UNIX	Centralized/SMP	Oracle	Sun	Hitachi
Anonymous	167,173	Sybase IQ	UNIX	Centralized/SMP	Sybase	Sun	EMC
KT IT-Group	136,641	DB2	UNIX	Centralized/Cluster	IBM	IBM	Hitachi
Anonymous	134,880	Sybase IQ	UNIX	Centralized/SMP	Sybase	HP	HP
China Telecom Corporation Co.,Ltd. GuangZhou Research Institute	133,924	Sybase IQ	UNIX	Centralized/SMP	Sybase	Sun	Sun

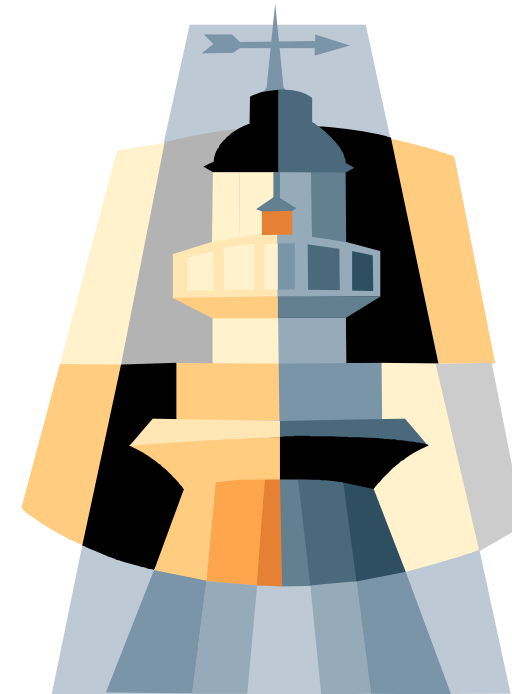
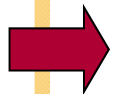
- http://www.wintercorp.com/VLDB/2005_TopTen_Survey/TopTenWinners_2005.asp

Felix Nau

Number of Rows is the total number of rows in all tables in the database, in millions.

10

- Einsatzgebiete
- OLAP versus OLTP
- Multidimensionale Modellierung
- OLAP Operationen
- Relationale Implementierung
- Neue Architekturen



Folien zu DWH: Ulf Leser (HU Berlin)

OLTP Beispiel

11

Login

```
SELECT pw FROM kunde WHERE login=„...“
UPDATE kunde SET last_acc=date, tries=0 WHERE
```

COMMIT

Willkommen

```
SELECT k_id, name FROM kunde WHERE login=„...“
SELECT last_pur FROM purchase WHERE k_id=...
```

COMMIT

Bestellung

```
SELECT av_qty FROM stock WHERE p_id=...
UPDATE stock SET av_qty=av_qty-1 where ...
INSERT INTO shop_cart VALUES( o_id, k_id, ...
```

COMMIT

Best. löschen

```
DELETE FROM shop_cart WHERE o_id=...
UPDATE stock SET av_qty=av_qty+1 where ...
```

COMMIT

OLAP Beispiele

12

- Welche Produkte hatten im letzten Jahr im Bereich Potsdam einen Umsatzrückgang um mehr als 10%?
 - Welche Produktgruppen sind davon betroffen?
 - Welche Lieferanten haben diese Produkte?
- Welche Kunden haben über die letzten 5 Jahre eine Bestellung über 50 Euro innerhalb von 4 Wochen nach einem persönlichen Anschreiben aufgegeben?
 - Wie hoch waren die Bestellungen im Durchschnitt?
 - Wie hoch waren die Bestellungen im Vergleich zu den durchschnittlichen Bestellungen des jeweiligen Kunden in einem vergleichbaren Zeitraum?
 - Lohnen sich Mailing-Aktionen?
- Haben solche Zweigstellen einen höheren Umsatz, die gemeinsam gekaufte Produkte nebeneinander platzieren?
 - Welche Produkte werden überhaupt zusammen gekauft – und wo?

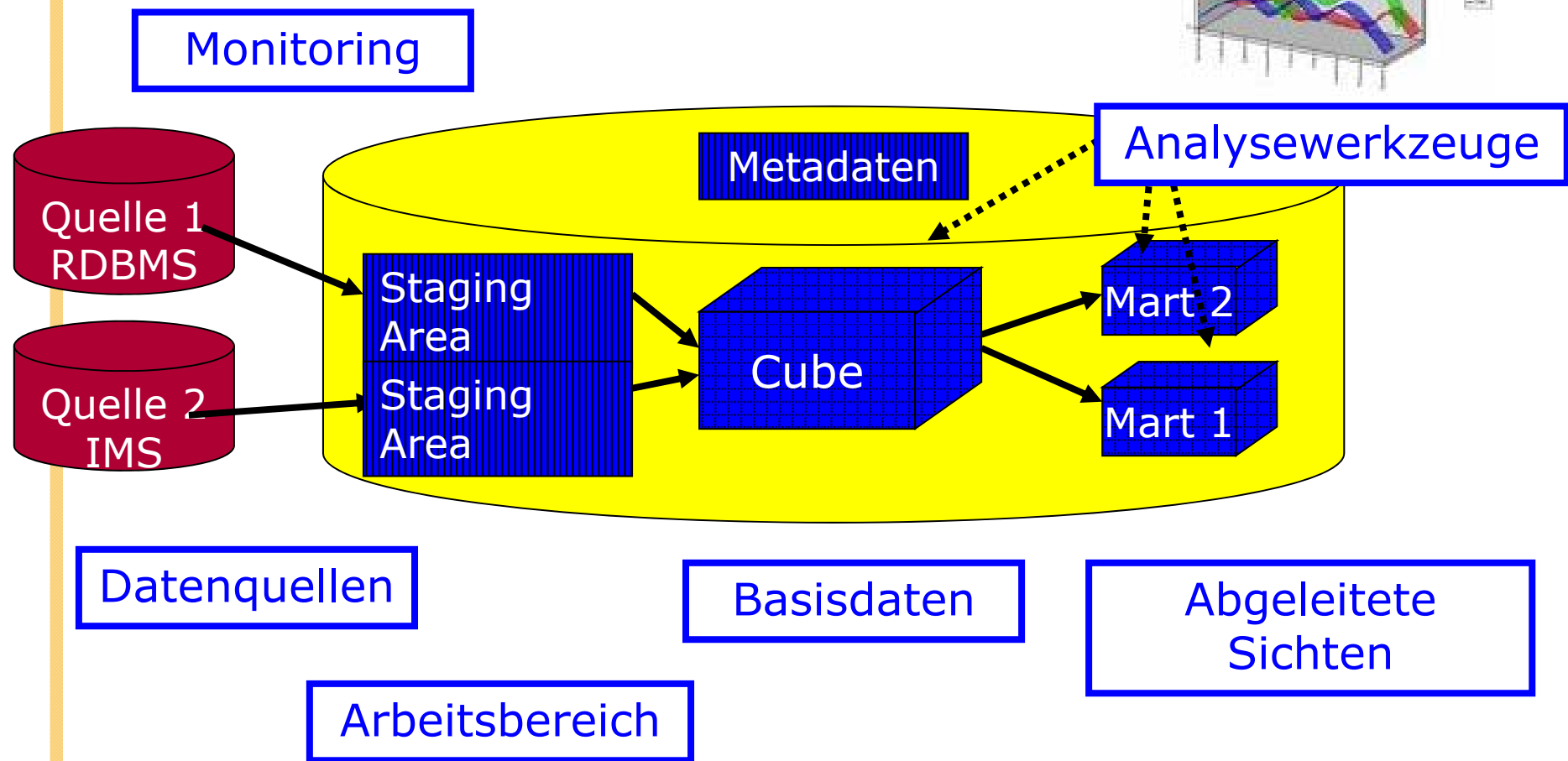
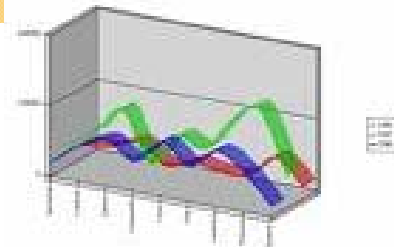
OLAP versus OLTP

13

	OLTP	OLAP
Typische Operationen	Insert, Update, Delete, Select	Select, Bulk-Inserts
Transaktionen	viele, kurze	Lange Lesetransaktionen
Typische Anfragen	Einfache Anfragen, Primärschlüsselzugriff, Schnelle Abfolgen von Selects/inserts/updates/deletes	Komplexe Anfragen: Aggregate, Gruppierung, Subselects, etc. Bereichsanfragen über mehrere Attribute
Daten pro Operation	Wenige Tupel	Mega-/ Gigabyte
Datenmenge in DB	Gigabyte	Terabyte
Eigenschaften der Daten	Rohdaten, häufige Änderungen	Abgeleitete Daten, historisch & stabil
Erwartete Antwortzeiten	msec - sec	Minuten - Stunden
Modellierung	Anwendungsorientiert	Themenorientiert
Typische Benutzer	Sachbearbeiter, Kunde	Management

DWH Architektur & Komponenten

14

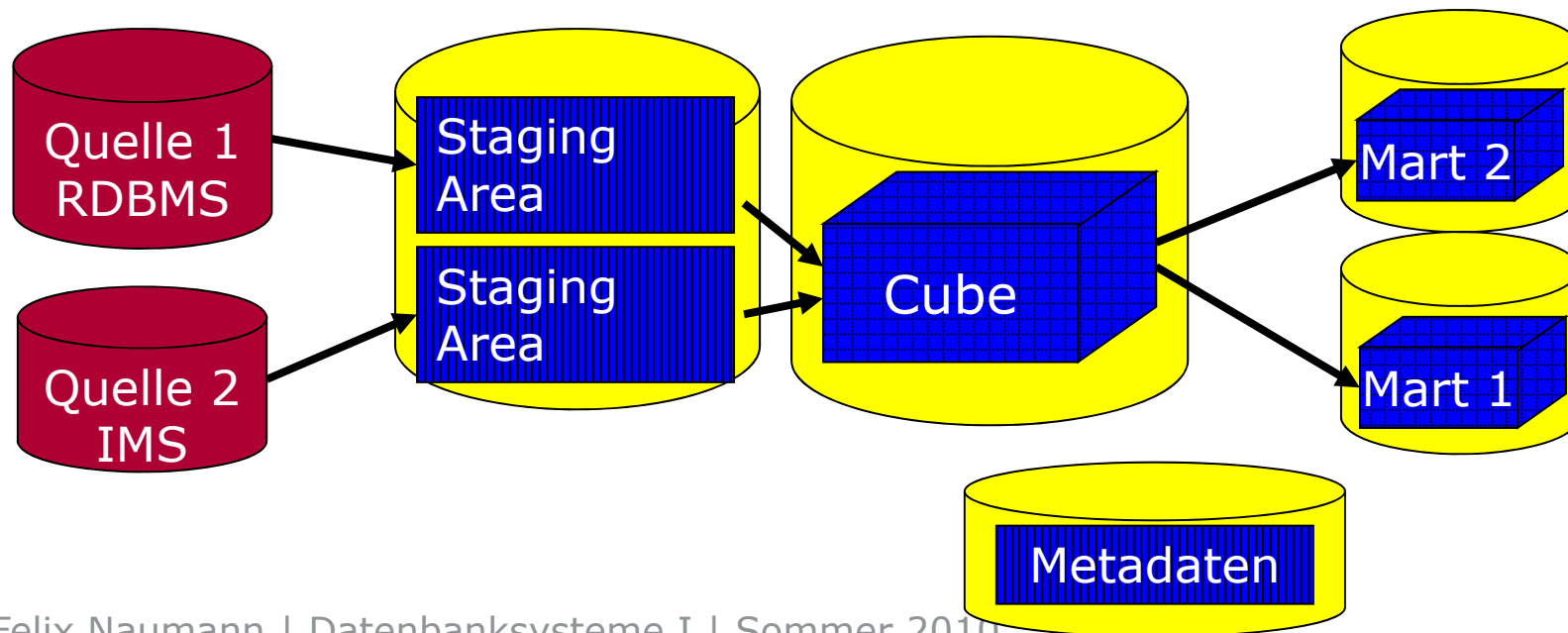


Alternativen

15

Physikalische Aufteilung variabel

- Data Marts auf eigenen Rechnern (Laptop)
- Staging Area auf eigenen Servern
- Metadaten auf eigenem Server (Repository)



Arbeitsbereich

16

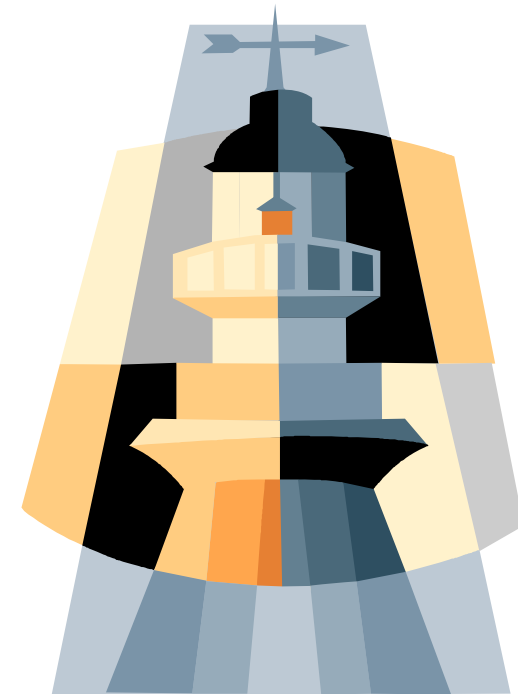
- Staging Area
 - Temporärer Speicher
 - Quellnahes Schema
- Motivation
 - ETL Arbeitsschritte effizienter implementierbar
 - ◇ Mengenoperationen, SQL
 - Zugriff auf Basisdatenbank möglich (Lookups)
 - Vergleich zwischen Datenquellen möglich
 - Filterfunktion: Nur einwandfreie Daten in Basisdatenbank übernehmen

Basisdatenbank

17

- Zentrale Komponente des DWH
 - Begriff „DWH“ meint oft nur die Basisdatenbank.
- Speichert Daten in feinsten Auflösung
 - Einzelne Verkäufe
 - Einzelne Bons
- Historische Daten
- Große Datenmengen
 - Spezielle Modellierung
 - Spezielle Optimierungsstrategien

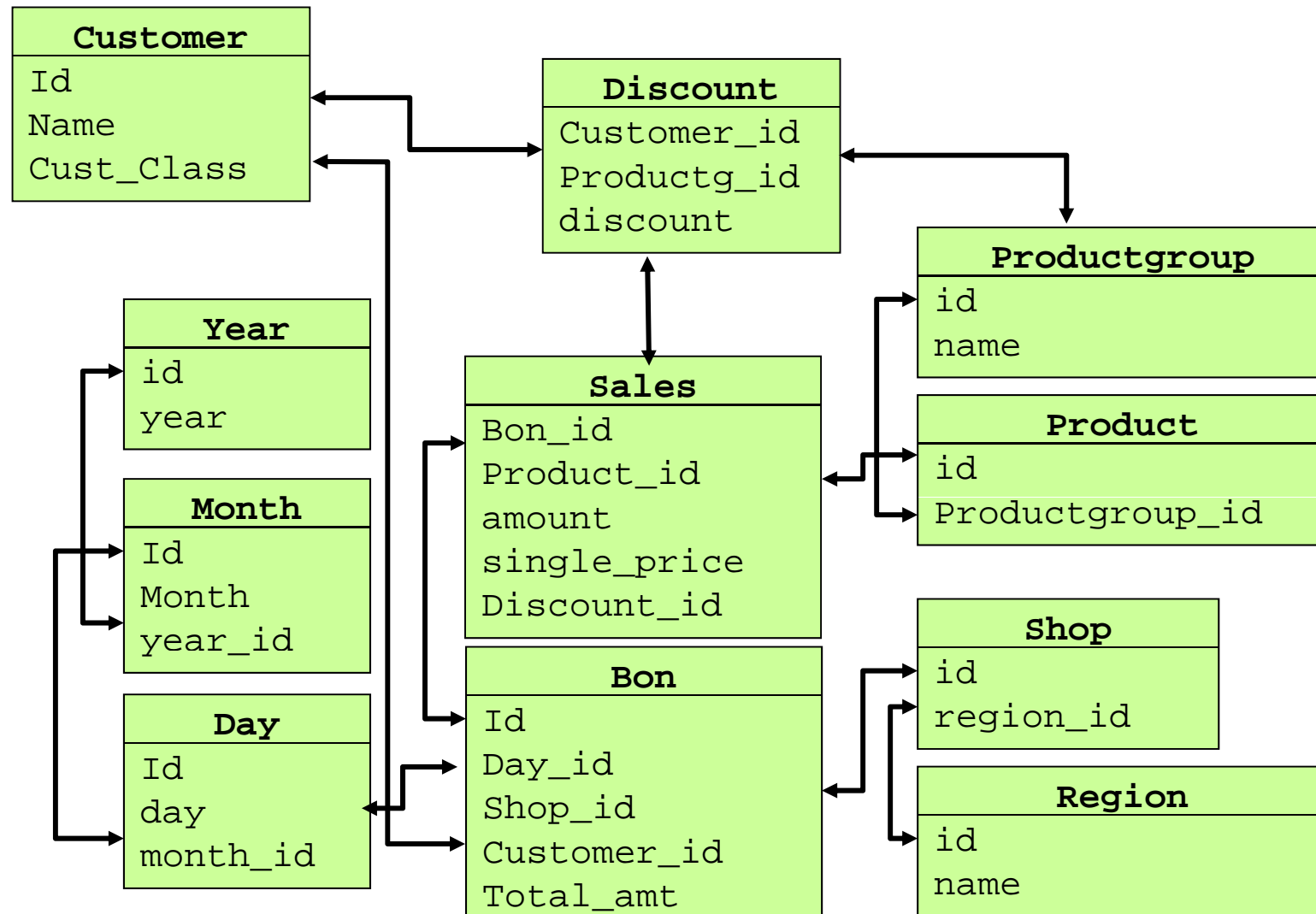
- Einsatzgebiete
- OLAP versus OLTP
- ➔ ■ Multidimensionale Modellierung
- OLAP Operationen
- Relationale Implementierung
- Neue Architekturen



Folien zu DWH: Ulf Leser (HU Berlin)

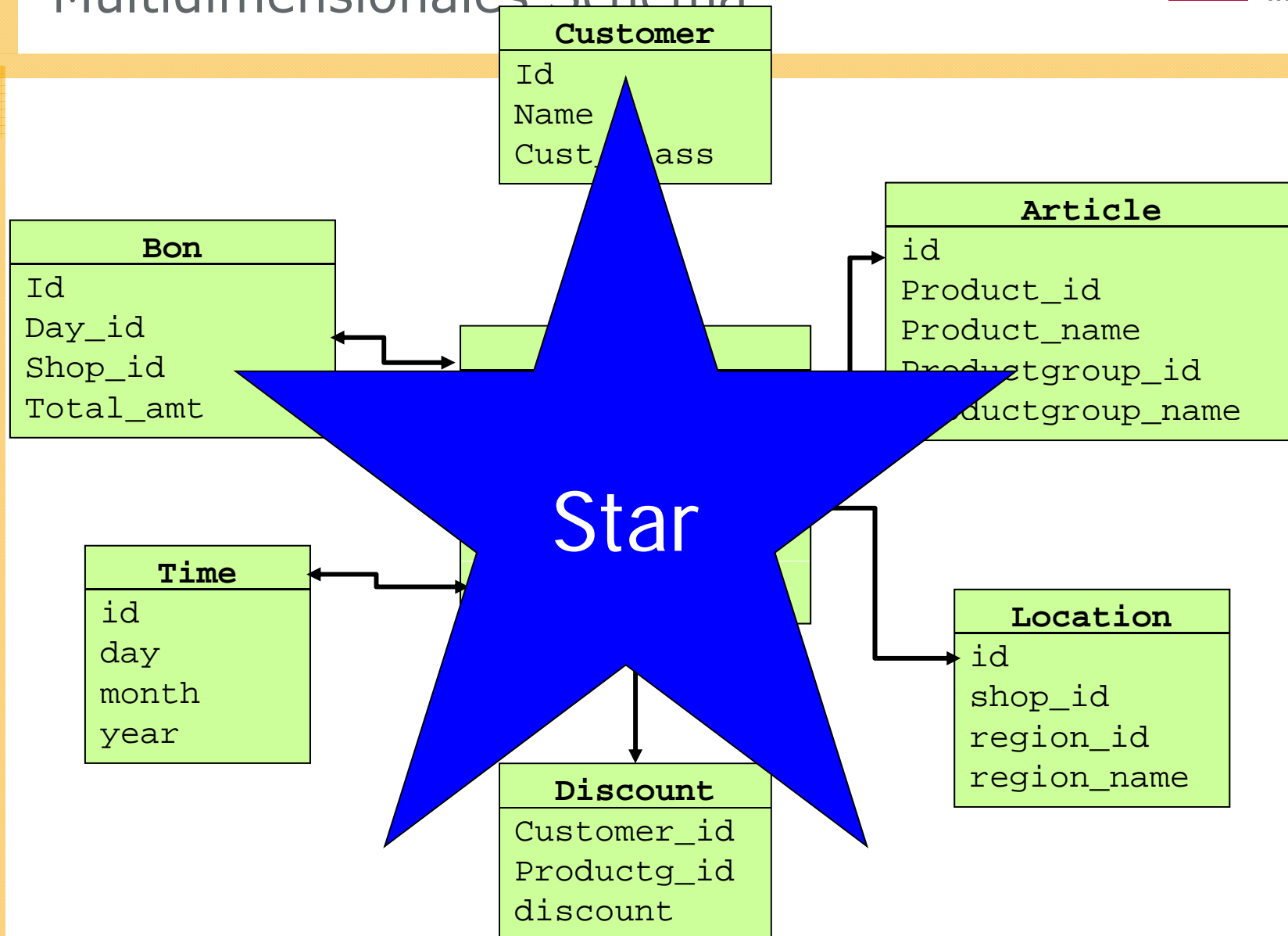
Beispiel: Normalisiertes Schema

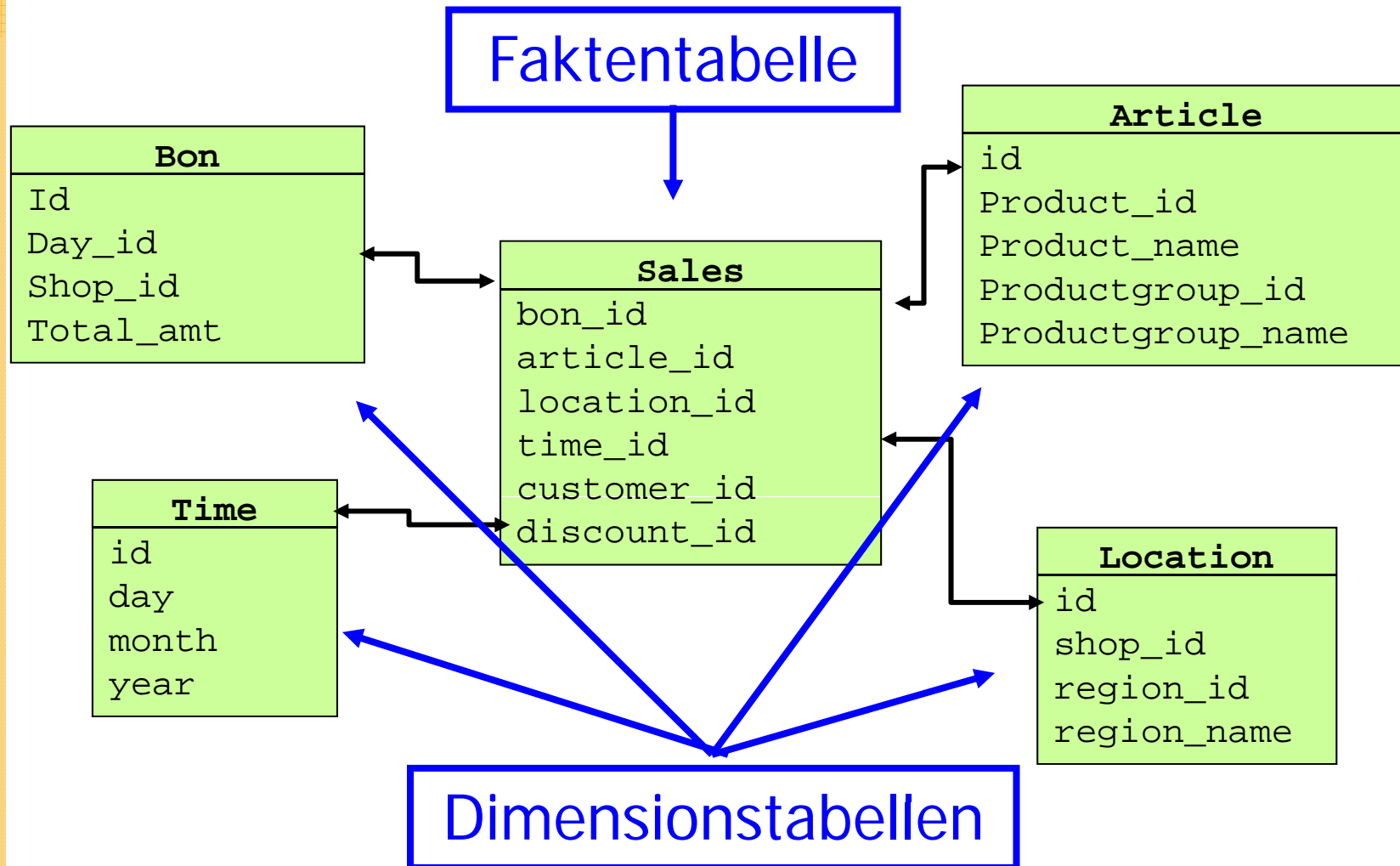
19



Multidimensionales Schema

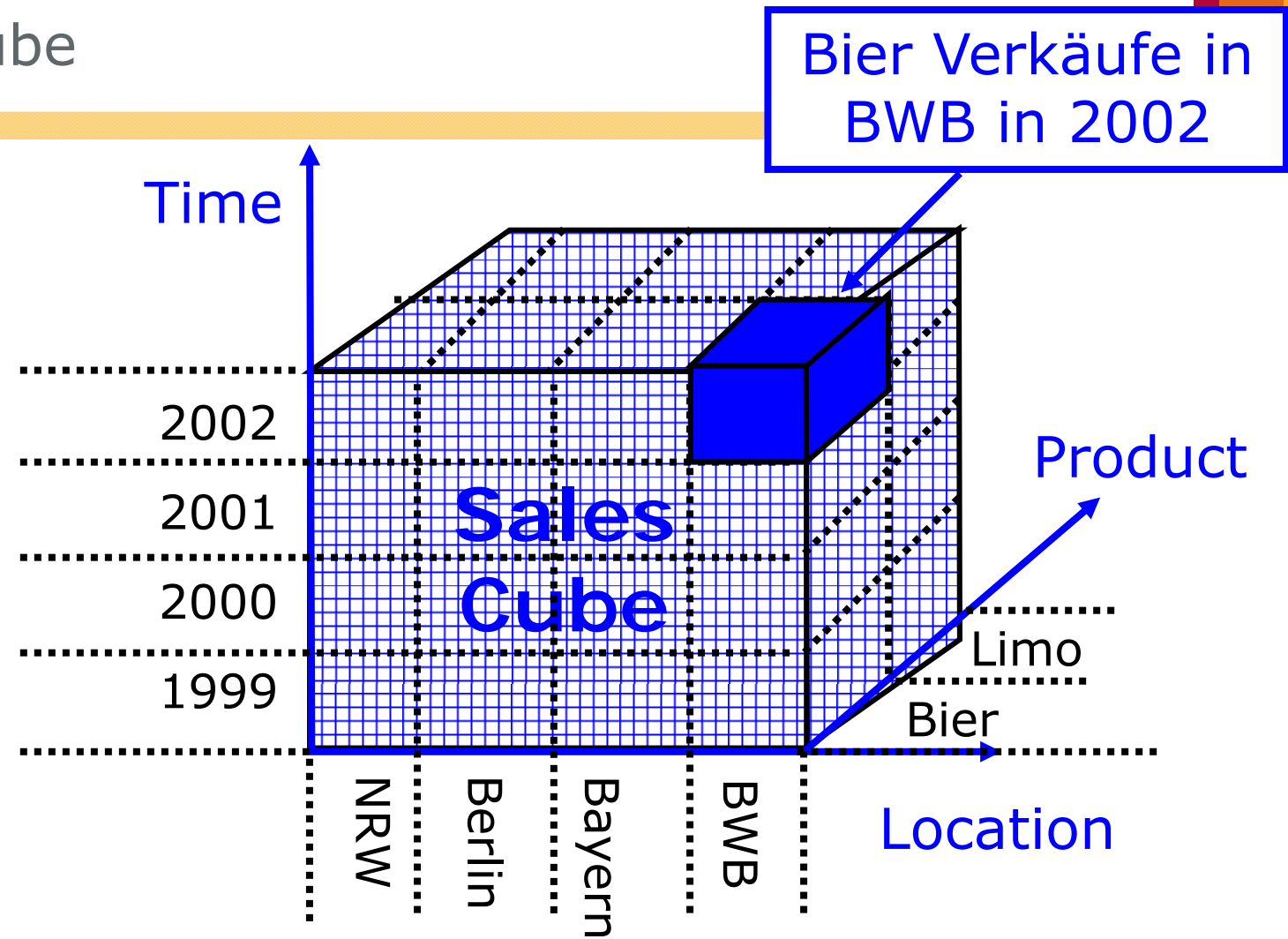
20





Cube

22



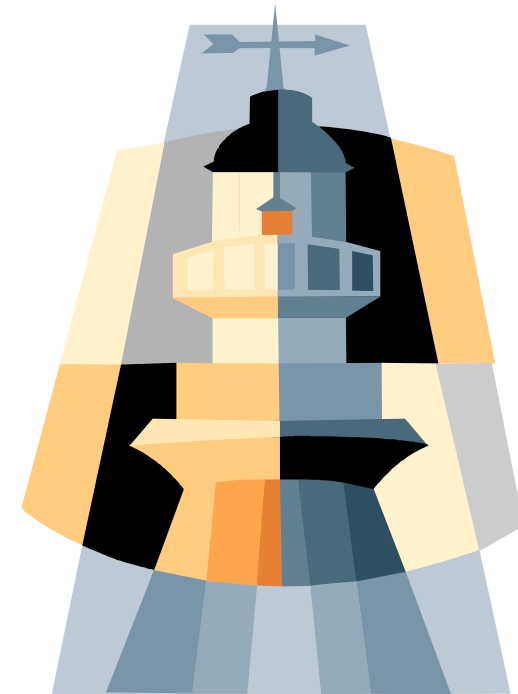
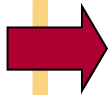
Cube -> Hypercube: Bon / Lieferant / Kunde / ...

Dimensionen

23

- Eindeutige Strukturierung des Datenraums
- Jede Dimension hat ein Schema
 - Tag, Woche, Jahr
 - Landkreis, Land, Staat
 - Produkt, Produktgruppe, Produktklasse, Produktfamilie
- ... und Wertebereiche
 - (1, 2, 3, ..., 31), (1, ... 52), (1900, ..., 2003)
 - (...), (Berlin, NRW, Department-1, ...), (BRD, F, ...)

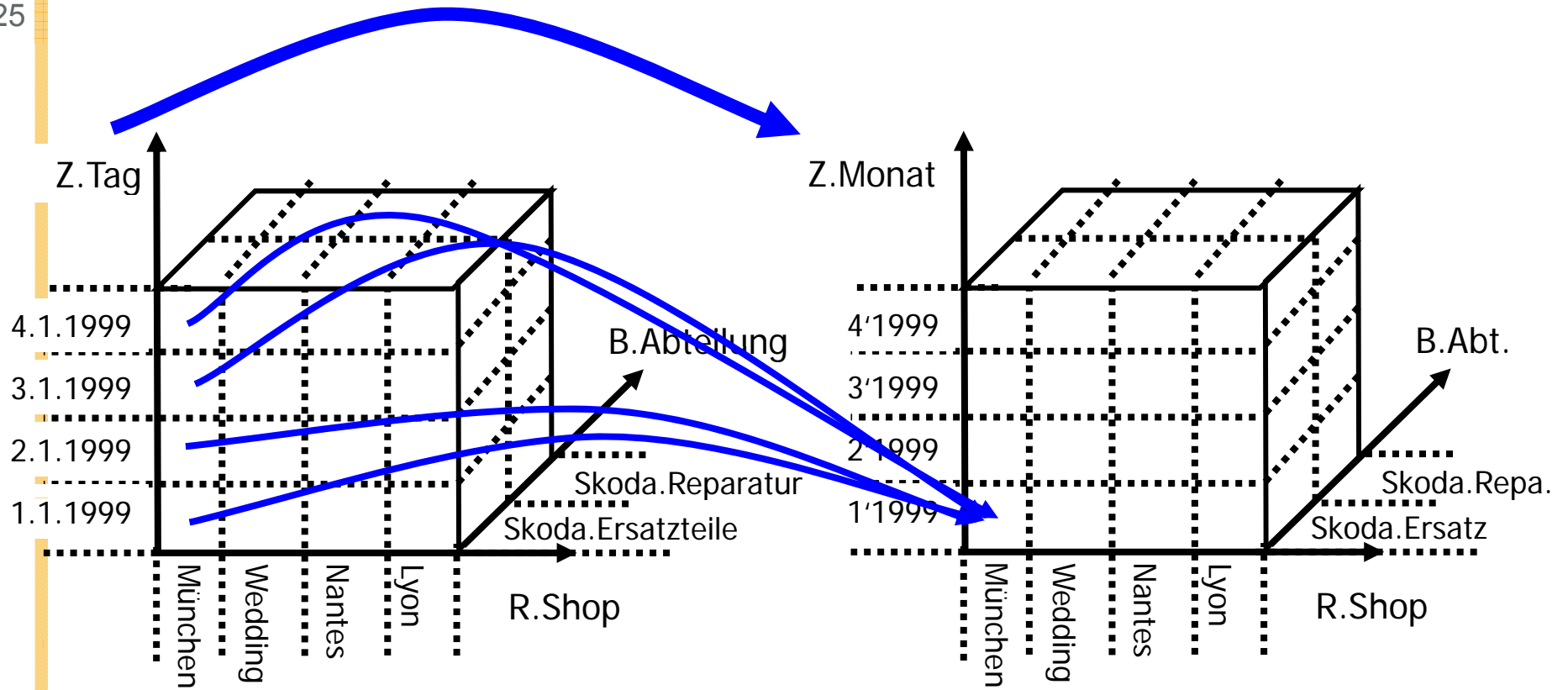
- Einsatzgebiete
- OLAP versus OLTP
- Multidimensionale Modellierung
- OLAP Operationen
- Relationale Implementierung
- Neue Architekturen



Folien zu DWH: Ulf Leser (HU Berlin)

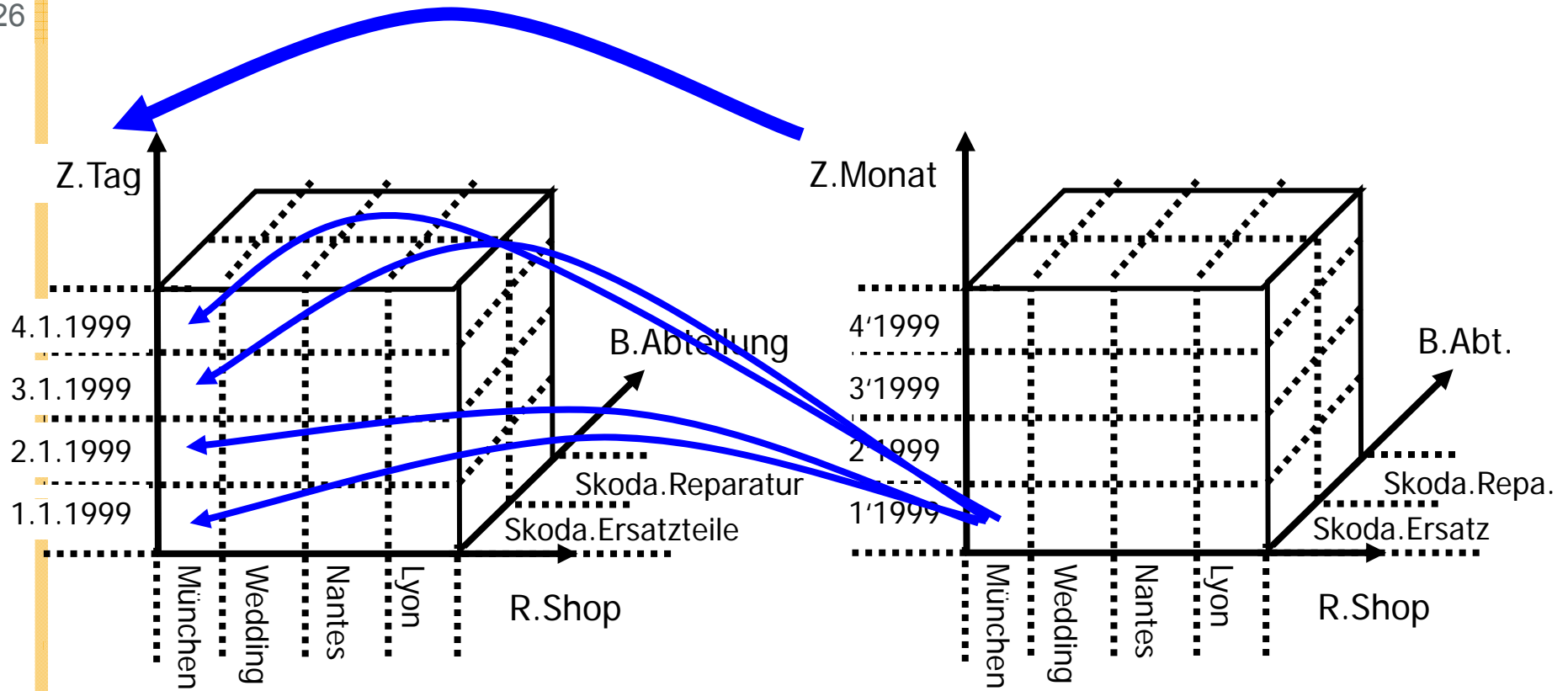
Beispiel: Aggregation (Roll-Up)

25



Beispiel: Verfeinerung (Drill-Down)

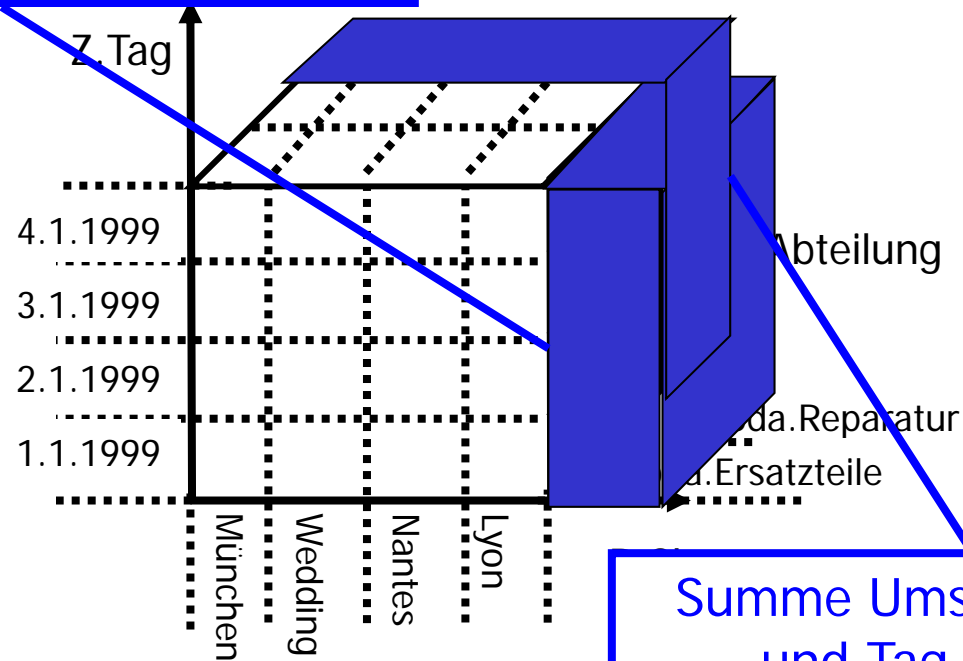
26



Aggregation bis TOP

27

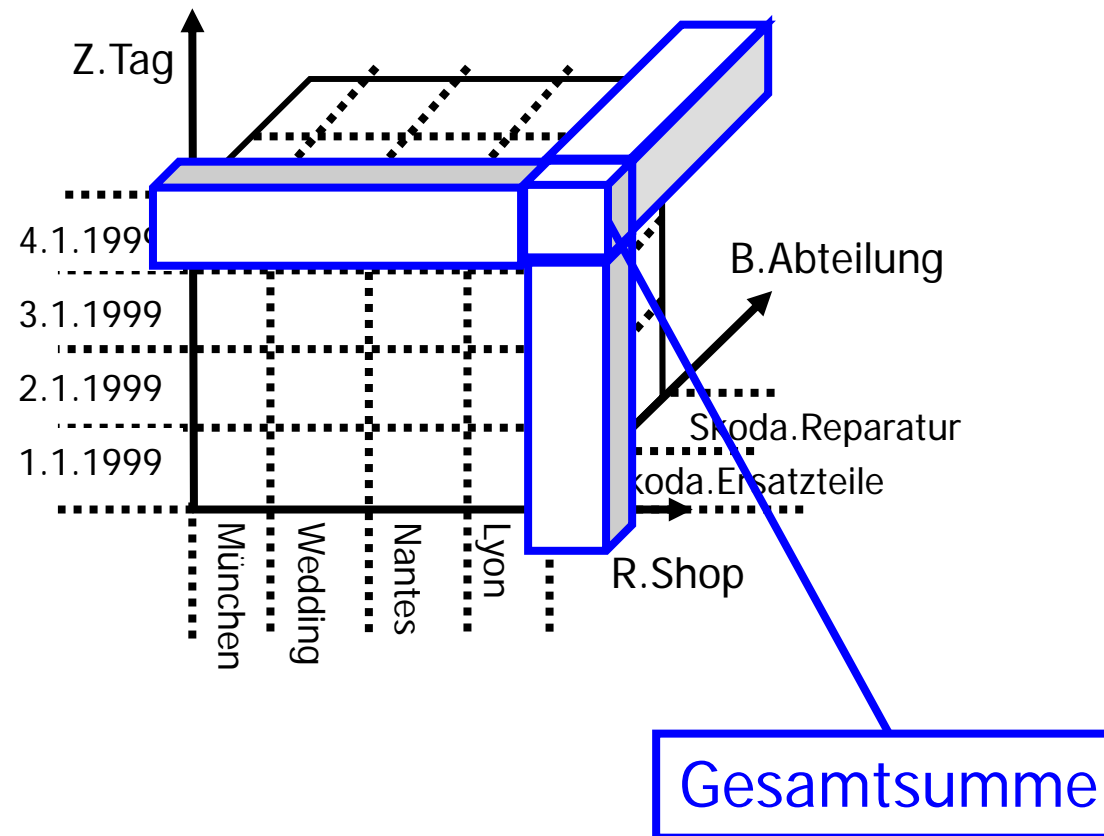
Summe Umsatz pro Tag und Abteilungen über alle Shops



Summe Umsatz pro Shop und Tag, über alle Abteilungen

... in mehreren Dimensionen

28

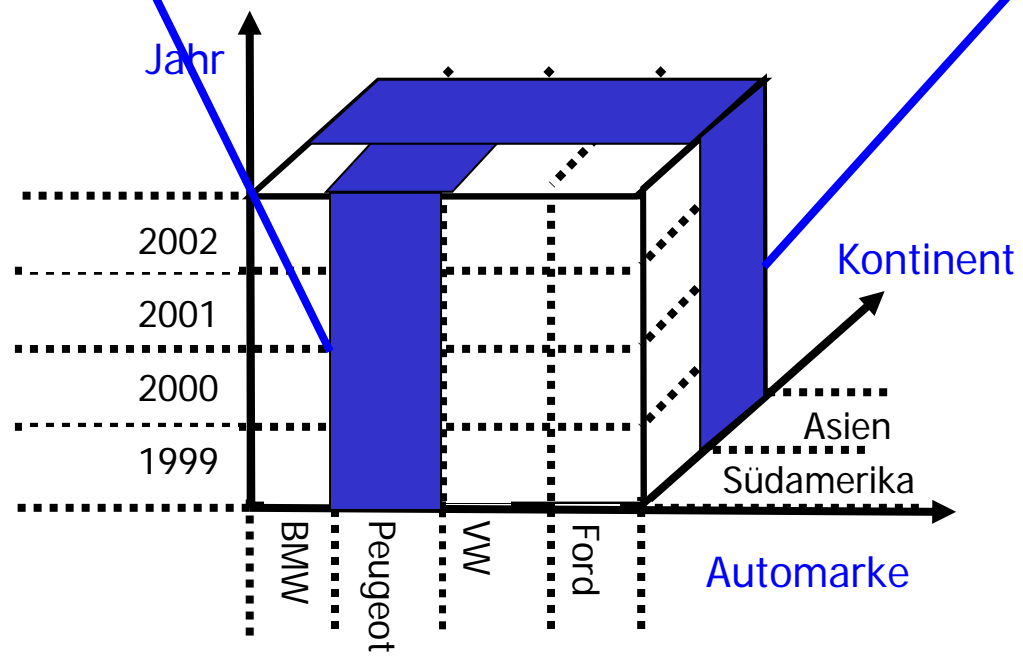


Selektion einer Scheibe (Slicing)

29

Verkäufe von Peugeot pro Jahr und Kontinent

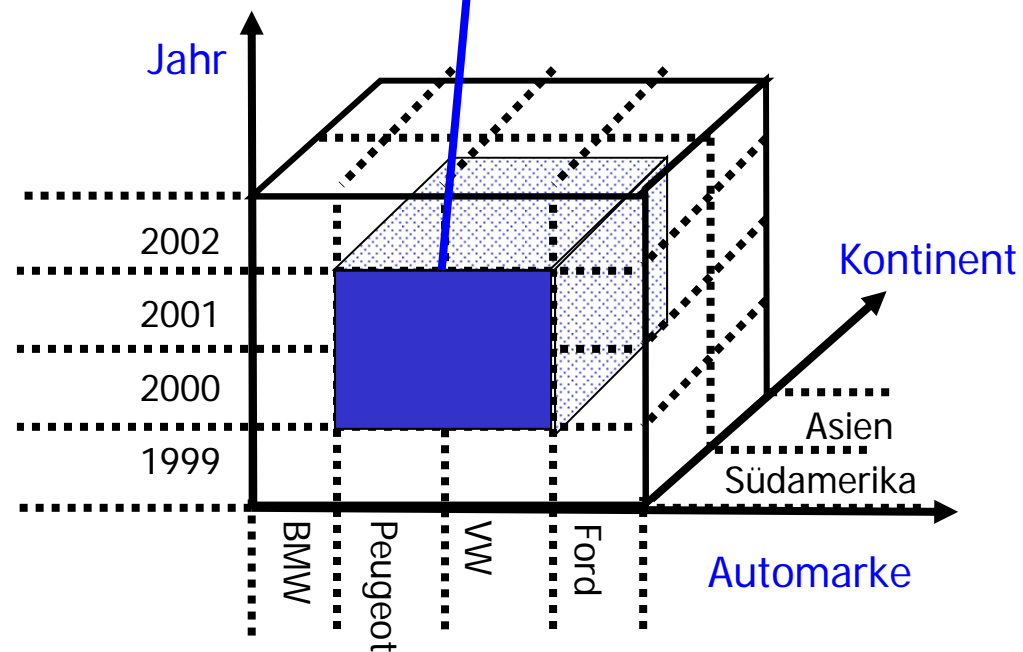
Verkäufe in Asien pro Jahr und Marke



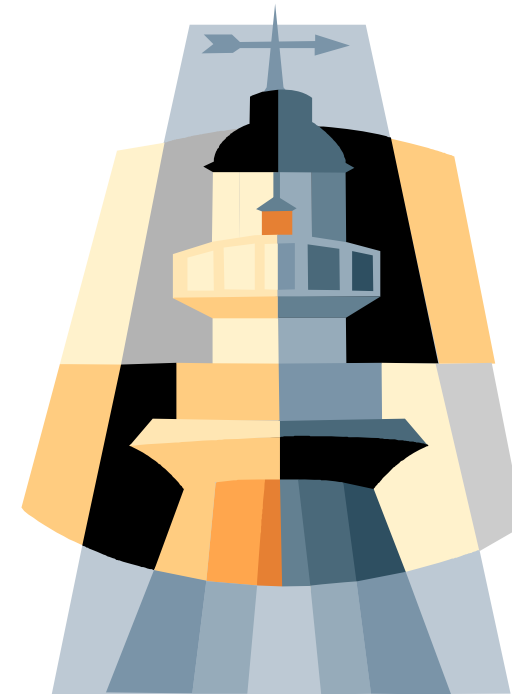
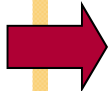
Auswahl von Unterwürfeln (*Dicing*)

30

Verkäufe von (Peugeot, VW) in (2000, 2001) pro Kontinent



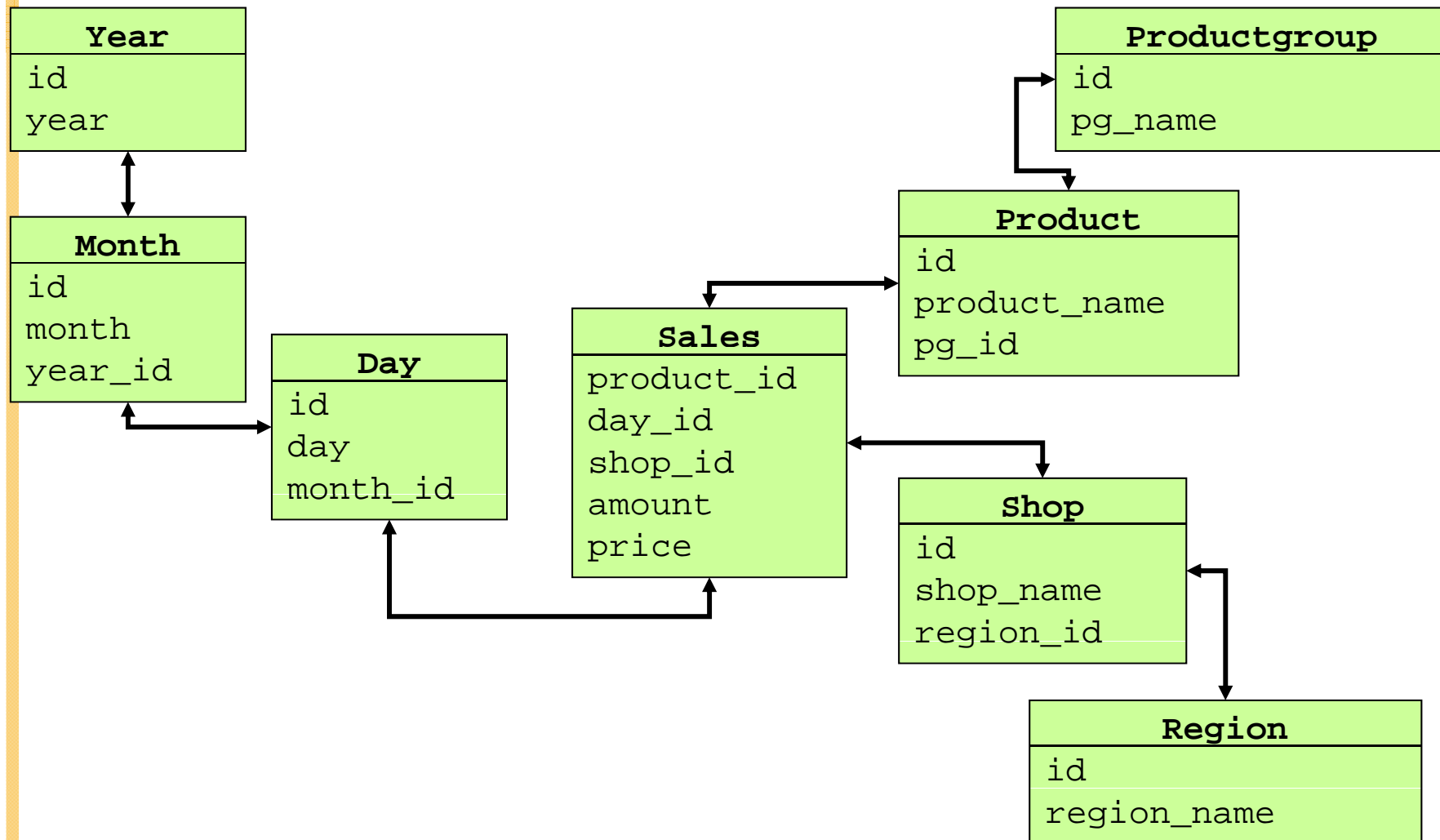
- Einsatzgebiete
- OLAP versus OLTP
- Multidimensionale Modellierung
- OLAP Operationen
- Relationale Implementierung
- Neue Architekturen



Folien zu DWH: Ulf Leser (HU Berlin)

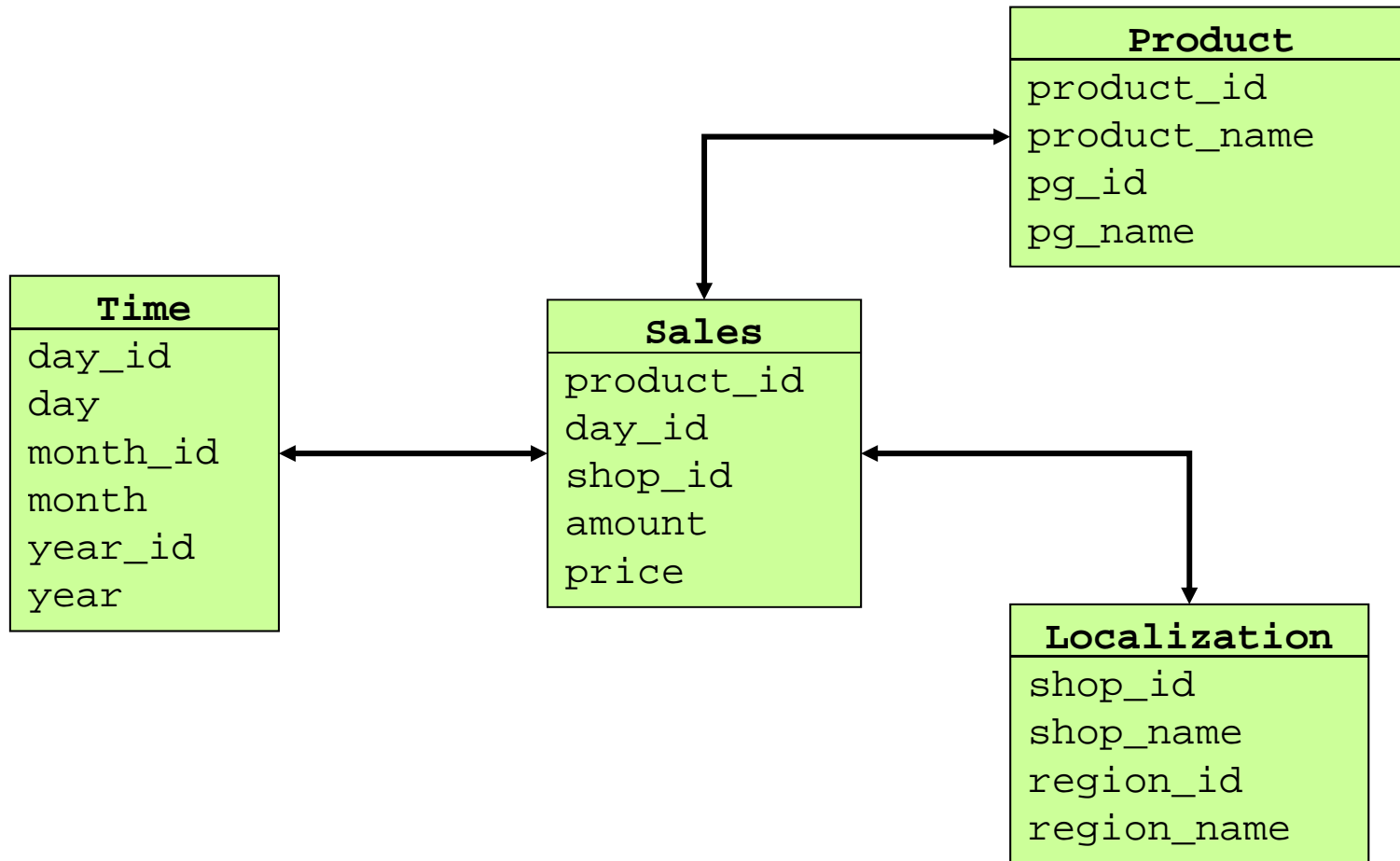
Variante 1 - Snowflake

32



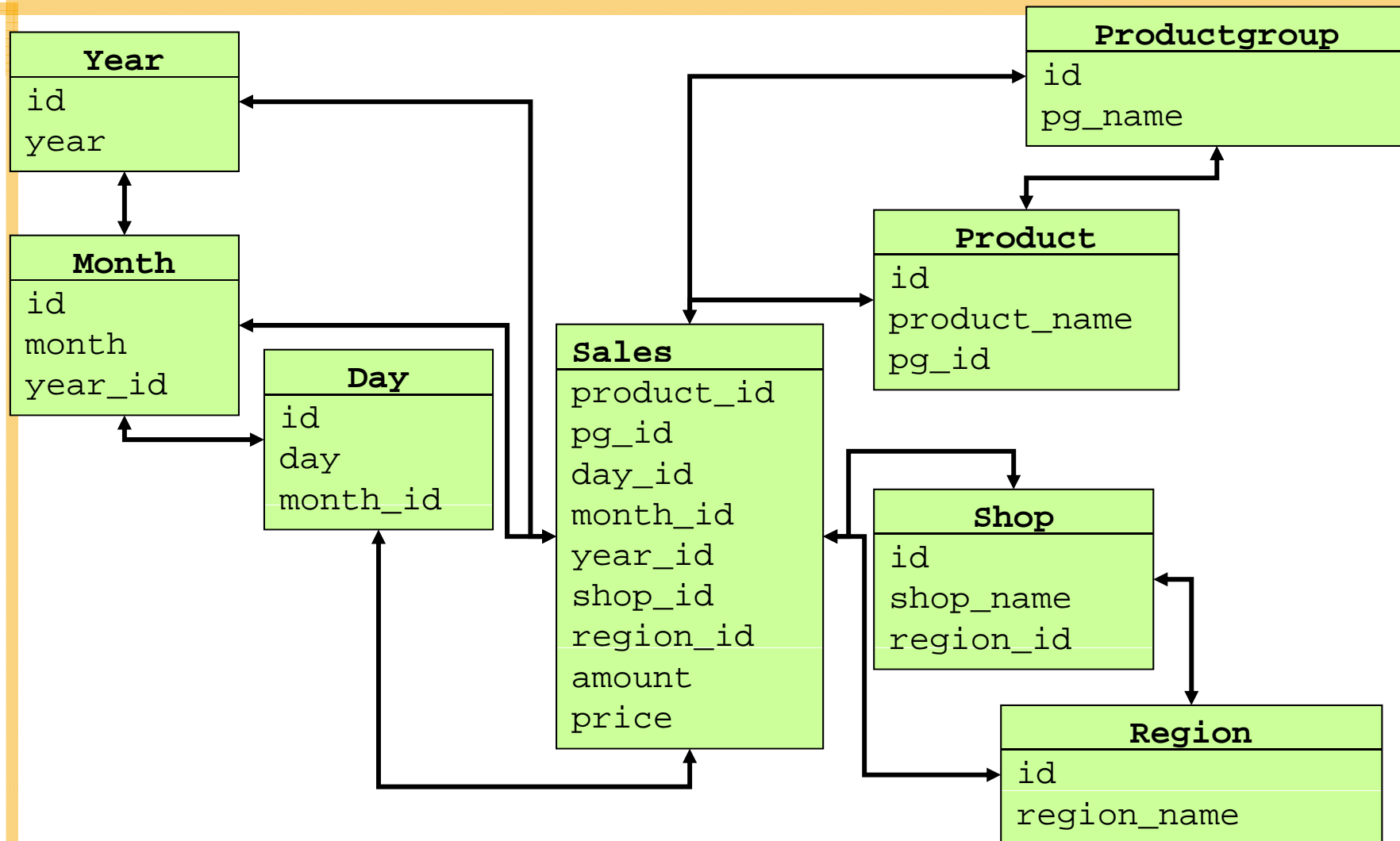
Variante 2: Star Schema

33



Variante 3: Fullfact

34



Fazit – Speicher und Query

35

- Speicherverbrauch Snowflake / Star praktisch identisch
 - Wenn Bedarf für Dimensionen vernachlässigbar
- Fullfact mit deutlich höherem Speicherverbrauch
 - Faktentabelle wird breiter
 - Dafür minimale Anzahl Joins
- Anzahl Joins: FullFact < Star < Snowflake
- Laufzeitverhalten hängt von mehr Faktoren als dem Schema ab
 - Bereichs- oder Punktanfrage
 - Indexierung
 - Selektivität der Bedingungen
 - Gruppierung und Aggregation
 - ...
- ... aber Joins sind tendenziell teuer

Roll-Up Operationen: Hierarchische Aggregation

36

- Wunsch: Verkaufsumsatz der Produktgruppe „Wein“ nach Tagen, Monaten und Jahren
- ```
SELECT T.year_id, T.month_id, T.day_id,
 sum(s.amount)
FROM Sales S, Product P, Time T
WHERE P.pg_name=„Wein“
AND P.product_id = S.product_id
AND T.day_id = S.day_id
GROUP BY T.year_id, T.month_id, T.day_id
```
- Summe nur für Tage (unterteilt nach Monaten/Jahren)
- Keine Summen pro Monat / pro Jahr
- Wunsch nicht in einer Anfrage formulierbar

|      |     |     |     |
|------|-----|-----|-----|
| 1997 | 1   | 1   | 150 |
| 1997 | 1   | 2   | 130 |
| 1997 | 1   | 3   | 145 |
| 1997 | 1   | 4   | 122 |
| ...  | ... | ... | ... |
| 1997 | 1   | 31  | 145 |
| 1997 | 2   | 1   | 133 |
| 1997 | 2   | 2   | 122 |
| ...  | ... | ... | ... |
| 1997 | 3   | 10  | 180 |
| 1997 | 12  | 31  | 480 |
| 1998 | 1   | 1   | 240 |
| ...  | ... | ... | ... |
| 2003 | 6   | 18  | 345 |

# Hierarchische Aggregation -2-

37

- Alle Verkäufe der Produktgruppe „Wein“ nach Tagen, Monaten und Jahren
- Benötigt UNION und eine Anfrage pro Klassifikationsstufe

```
SELECT T.day_id, sum(amount*price)
FROM Sales S, Product P
WHERE P.pg_name=„Wein“ and
```

```
SELECT T.month_id, sum(amount*price)
FROM Sales S, Product P, Time T
WHERE P.pg_name=„Wein“ and
```

```
SELECT T.year_id, sum(amount*price)
FROM Sales S, Product P, Time T
WHERE P.pg_name=„Wein“ and
P.product_id = S.product_id and
T.day_id = S.day_id
GROUP BY T.year_id
```

# ROLLUP Operator

38

- Herkömmliches SQL
  - Dimension mit k Stufen – Union von k Queries
  - k Scans der Faktentabelle
    - ◇ Keine Optimierung wg. fehlender Multiple-Query Optimierung in kommerziellen RDBMS
  - Schlechte Ergebnisreihenfolge
  
- ROLLUP Operator
  - Hierarchische Aggregation mit Zwischensummen
  - Summen werden durch „ALL“ als Wert repräsentiert

# ROLLUP Beispiel

39

```
SELECT T.year_id, T.month_id, T.day_id, sum(...)
FROM Sales S, Time T
WHERE T.day_id = S.day_id
GROUP BY ROLLUP(T.year_id, T.month_id, T.day_id)
```

|      |       |     |            |
|------|-------|-----|------------|
| 1997 | Jan   | 1   | 200        |
| 1997 | Jan   | ... |            |
| 1997 | Jan   | 31  | 300        |
| 1997 | Jan   | ALL | 31.000     |
| 1997 | Feb   | ... |            |
| 1997 | March | ALL | 45.000     |
| 1997 | ...   | ... |            |
| 1997 | ALL   | ALL | 1.456.400  |
| 1998 | Jan   | 1   | 100        |
| 1998 | ...   | ... |            |
| 1998 | ALL   | ALL | 2.450.000  |
| ...  | ...   | ... |            |
| ALL  | ALL   | ALL | 12.445.750 |

# Multidimensionale Aggregation

40

|        | 1998 | 1999 | 2000 | Gesamt |
|--------|------|------|------|--------|
| Weine  | 15   | 17   | 13   | 45     |
| Biere  | 10   | 15   | 11   | 36     |
| Gesamt | 25   | 32   | 24   | 81     |

- `sum( ) ... GROUP BY pg_id, year_id`
- `sum( ) ... GROUP BY pg_id`
- `sum( ) ... GROUP BY year_id`
- `sum( )`



# Cube Operator

41

- $d$  Dimensionen, jeweils eine Klassifikationsstufe
  - Jede Dimension kann in Gruppierung enthalten sein oder nicht
  - $2^d$  Gruppierungsmöglichkeiten
- Herkömmliches SQL
  - Viel Schreibarbeit
  - $2^d$  Scans der Faktentabelle (wieder keine Optimierung möglich)
- CUBE Operator
  - Berechnung der Summen von sämtlichen Kombinationen der Argumente (Klassifikationsstufen)
  - Summen werden durch „ALL“ repräsentiert

# Einfacher SQL Ansatz

42

```
SELECT Marke, Farbe, SUM(Verkäufe)
FROM AutoTab
GROUP BY (Marke, Farbe)
```

UNION

```
SELECT Marke, ALL, SUM(Verkäufe)
FROM AutoTab
GROUP BY (Marke)
```

UNION

```
SELECT ALL, Farbe, SUM(Verkäufe)
FROM AutoTab
GROUP BY (Farbe)
```

UNION

```
SELECT ALL, ALL, SUM(Verkäufe)
FROM AutoTab;
```

| Marke | Farbe | Verkäufe |
|-------|-------|----------|
| VW    | Blau  | 32       |
| VW    | Weiß  | 17       |
| VW    | Rot   | 5        |
| Opel  | Blau  | 24       |
| Opel  | Weiß  | 19       |
| Opel  | Rot   | 12       |
| VW    | ALL   | 54       |
| Opel  | ALL   | 55       |
| ALL   | Blau  | 56       |
| ALL   | Weiß  | 36       |
| ALL   | Rot   | 17       |
| ALL   | ALL   | 109      |

# Ansatz mit CUBE-Operator

43

## ■ Neuer Ansatz

- SELECT Marke, Farbe, SUM(Verkäufe)  
FROM AutoTab  
GROUP BY **CUBE(Marke, Farbe);**
- Unterschiede in der Syntax:
  - keine UNIONS mehr notwendig  
⇒ einfachere Anfrage
- Unterschiede in der Semantik:
  - Keine

## Bisheriger Ansatz

```
SELECT Marke, Farbe, SUM(Verkäufe)
FROM AutoTab
GROUP BY (Marke, Farbe)
```

UNION

```
SELECT Marke, ALL, SUM(Verkäufe)
FROM AutoTab
GROUP BY (Marke)
```

UNION

```
SELECT ALL, Farbe, SUM(Verkäufe)
FROM AutoTab
GROUP BY (Farbe)
```

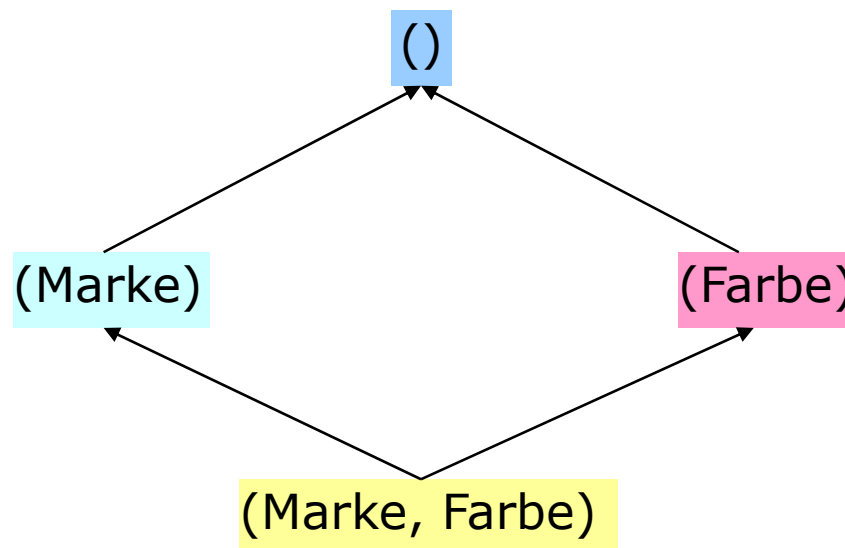
UNION

```
SELECT ALL, ALL, SUM(Verkäufe)
FROM AutoTab;
```

# Ansatz mit CUBE-Operator

44

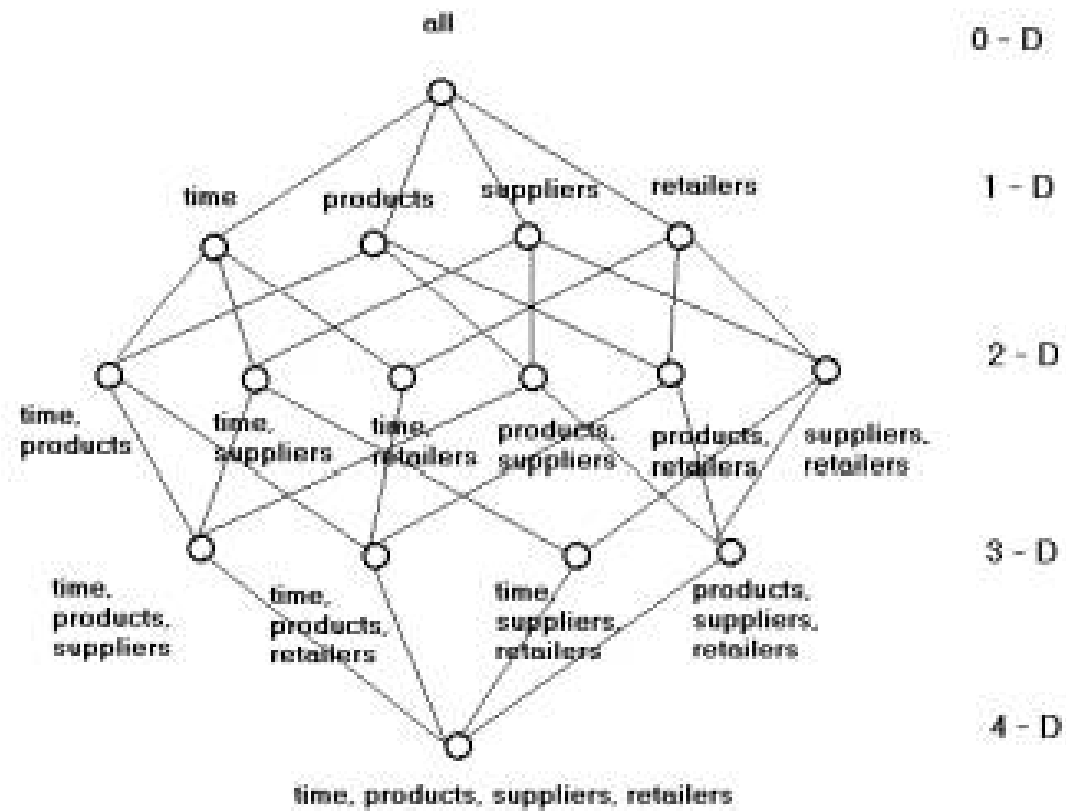
- **Ableitbarkeit der Gruppen**
- Beziehung lässt sich mithilfe eines Aggregationsgitters darstellen
- $(X,Y) \triangleq$  Gruppierung über X und Y



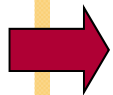
| Marke | Farbe | Verkäufe |
|-------|-------|----------|
| VW    | Blau  | 32       |
| VW    | Weiß  | 17       |
| VW    | Rot   | 5        |
| Opel  | Blau  | 24       |
| Opel  | Weiß  | 19       |
| Opel  | Rot   | 12       |
| VW    | ALL   | 54       |
| Opel  | ALL   | 55       |
| ALL   | Blau  | 56       |
| ALL   | Weiß  | 36       |
| ALL   | Rot   | 17       |
| ALL   | ALL   | 109      |

# Another Lattice

45



- Einsatzgebiete
- OLAP versus OLTP
- Multidimensionale Modellierung
- OLAP Operationen
- Relationale Implementierung
- Neue Architekturen
  - Folie nach:  
Hasso Plattner, *A Common Database Approach for OLTP and OLAP Using an In-Memory Column Database*  
*SIGMOD 2009*



# Row storage

47

