Data Warehousing

Weitere Buzzwörter:

OLAP, Decision Support, Data Mining

Wichtige Hinweise

- Zu diesem Thema gibt es eine Spezialvorlesung im Sommersemester
- Hier nur grober Überblick über Idee und einige Features – „Teaser“ für die Spezialvorlesung
- Literatur: Chaudhuri, Dayal: SIGMOD 1997

OLTP vs. OLAP

- OLTP – Online Transaction Processing
  - Viele kleine Transaktionen
    (Punktanfragen und UPDATE oder INSERT)
  - Möglichst wenig Redundanz, Normalisierte Schemas
  - Aktueller Datenbankzustand
- OLTP Beispiele:
  - Flugbuchung
  - Auftragsannahme, Rechnungswesen
    (Quelle, Telekom)
- Ziel: z.B. 6000 Transaktionen pro Sekunde

OLTP vs. OLAP

- OLAP – Online Analytical Processing
  - große, schwere Anfragen; keine Updates
  - Redundanz notwendig (Matrialisierte Ergebnisse, besondere Indexe, keine Normalisierung)
  - Tages- oder Wochenaktualität ausreichend
- OLAP Beispiele
  - Manager in einem Unternehmen
  - Statistisches Bundesamt (Volkszählung)
  - Analyse von Webtraces
- Ziel: Antwortzeit von wenigen Sekunden/Minuten

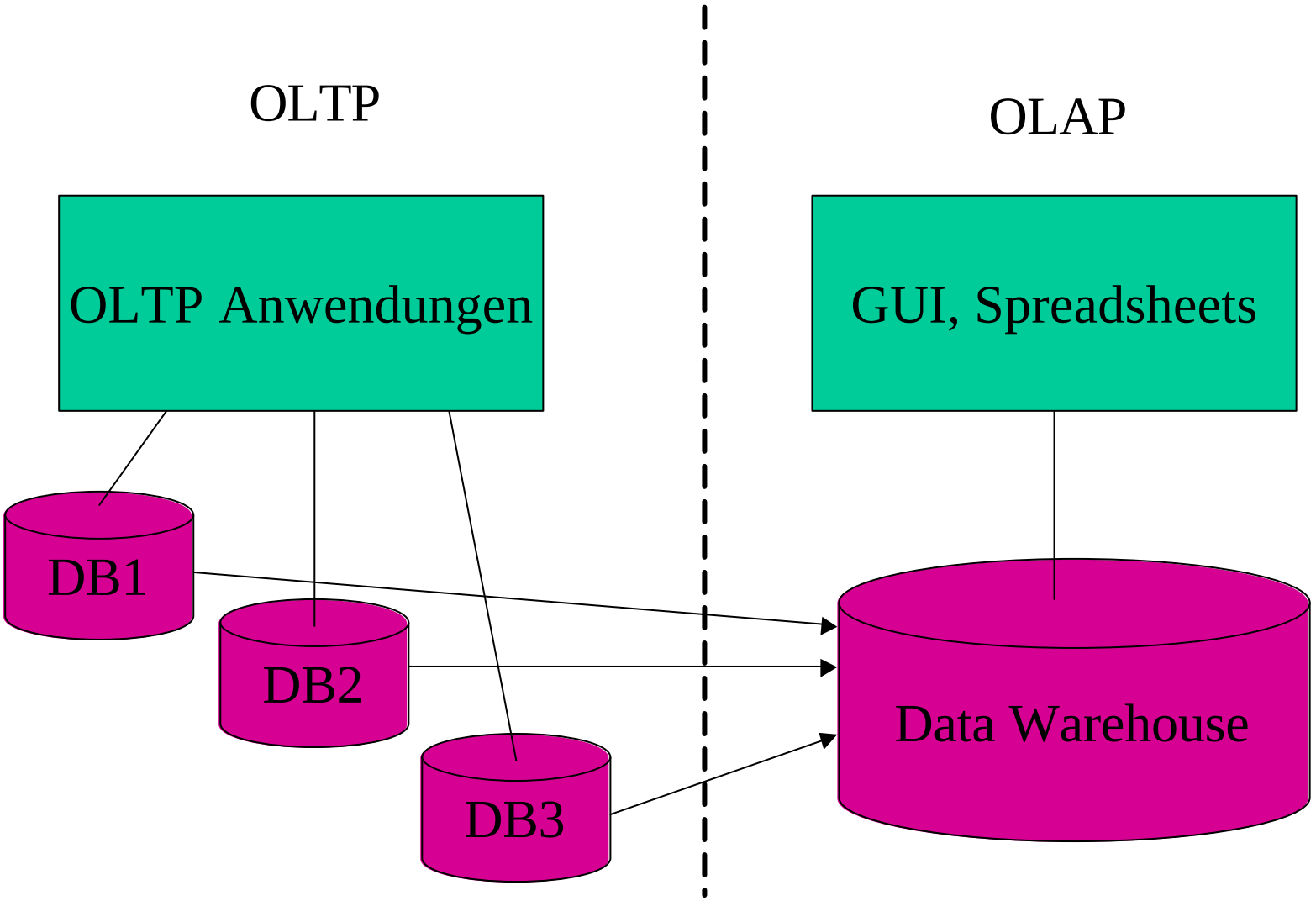
Konflikte

- Sperrkonflikte: OLAP behindert OLTP
- Datenbankentwurf:
  - OLTP normalisiert, OLAP nicht
- Tuning, Optimierereinstellungen
  - OLTP: inter-query Parallelität,
    möglichst kein Overhead bei Optimierung
  - OLAP: inter-query Parallelität,
    bestmöglicher Plan notwendig
- Aktualität der Daten: OLAP braucht oft reproduzierbare Ergebnisse
- Präzision: Sampling oft ausreichend für OLAP, OLAP arbeitet mit verdichteten Daten

Lösung: Data Warehouse

- Eigene Spielwiese für OLAP
- Daten werden in den OLTP Systemen gesammelt
- Daten werden im Warehouse repliziert (verdichtet und in einem besonderen Layout)
- Neue Daten werden periodisch ans Warehouse propagiert
- Veraltete Daten werden im Warehouse gelöscht und von den OLTP Systemen archiviert

Architektur



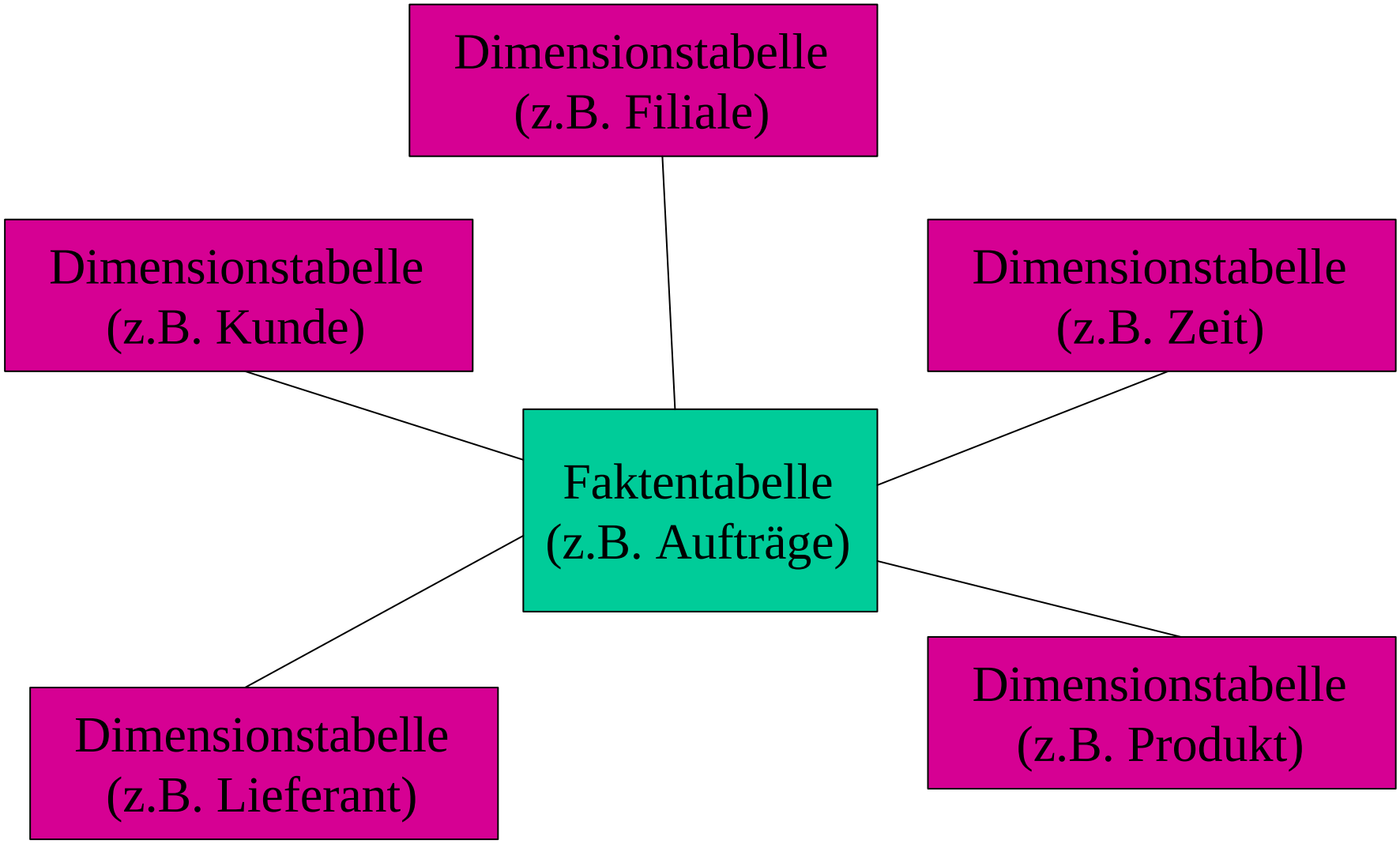
Data Warehouse in der Praxis

- Ersten industriellen Projekte seit ca. 1995
- In den ersten Jahren sind 80 Prozent der Projekte gescheitert
- Inzwischen gibt es spezielle Produkte und bessere Projektplanung und Beratung
- Risikofaktoren: Datenintegration, schlechte Abbildung des Geschäftsmodells
- Kosten sind allerdings nach wie vor hoch (Warehouse kostet soviel wie OLTP System)
- Großer Wachstumsmarkt – jeder macht es angeblich spart WalMart 20 Prozent der Kosten durch besseres Management mit einem DW

Produkte und Tools

- Oracle 8i, IBM UDB, Microsoft SQL Server, ...
- Also jeder Datenbankanbieter
- SAP Business Information Warehouse
- MicroStrategy
- Armada von Data Mining Tools
  (z.B. IBM Intelligent Miner, Webmining Tools)
- Datenintegration, verteilte Datenbanken
  (IBM DataJoiner, IBM Replicator, CORBA, ...)

Sternschema



Faktentabelle (Aufträge)

Nr.	Kunde	Datum	...	Filiale	Preis	Anz	TAX
001	Heinz	13.5.	...	Mainz	500	5	7.0
002	Ute	17.6.	...	Köln	500	1	14.0
003	Heinz	21.6.	...	Köln	700	1	7.0
004	Heinz	4.10.	...	Mainz	400	7	7.0
005	Karin	4.10.	...	Mainz	800	3	0.0
006	Thea	7.10.	...	Köln	300	2	14.0
007	Nobbi	13.11.	...	Köln	100	5	7.0
008	Sarah	20.12	...	Köln	200	4	7.0

Faktentabelle

- Aufbau:
  - Schlüssel (z.B. Auftragsnummer)
  - Fremdschlüssel zu allen Dimensionstabellen
  - Kennzahlen (z.B. Preis, Anzahl, ...)
- Speichern *Bewegungsdaten*
- Sind in der Regel sehr groß und normalisiert

Dimensionstabelle (Filiale)

Bez.	Leiter	Stadt	Region	Land	Telefon
Mainz	Helga	Mainz	Süd	D	1422
Köln	Vera	Hürth	Süd	D	3311

- Ist nicht normalisiert: Stadt -> Region -> Land
  hierdurch spart man sich „Joins“ bei Anfragen
- Dimensionstabellen sind in der Regel sehr viel kleiner als Faktentabellen
- Dimensionstabellen speichern *Stammdaten*
- Attribute werden *Merkmale* genannt

Typische Anfragen

```
SELECT  d1.x, d2.y, d3.z, sum(f.z1), avg(f.z2)
FROM    Fakten f, Dim1 d1, Dim2 d2, Dim3 d3
WHERE  a < d1.feld < b  AND  d2.feld = c AND
                Joinprädikate
GROUP BY d1.x, d2.y, d3.z;
```

- Selektiere nach bestimmten Merkmalen
  z.B. Kunde ist weiblich
- Gruppiere nach Merkmalen
  z.B. Regionen oder Monaten oder Quartalen
- Wende Aggregatfunktion auf Kennzahlen an

Beispiel

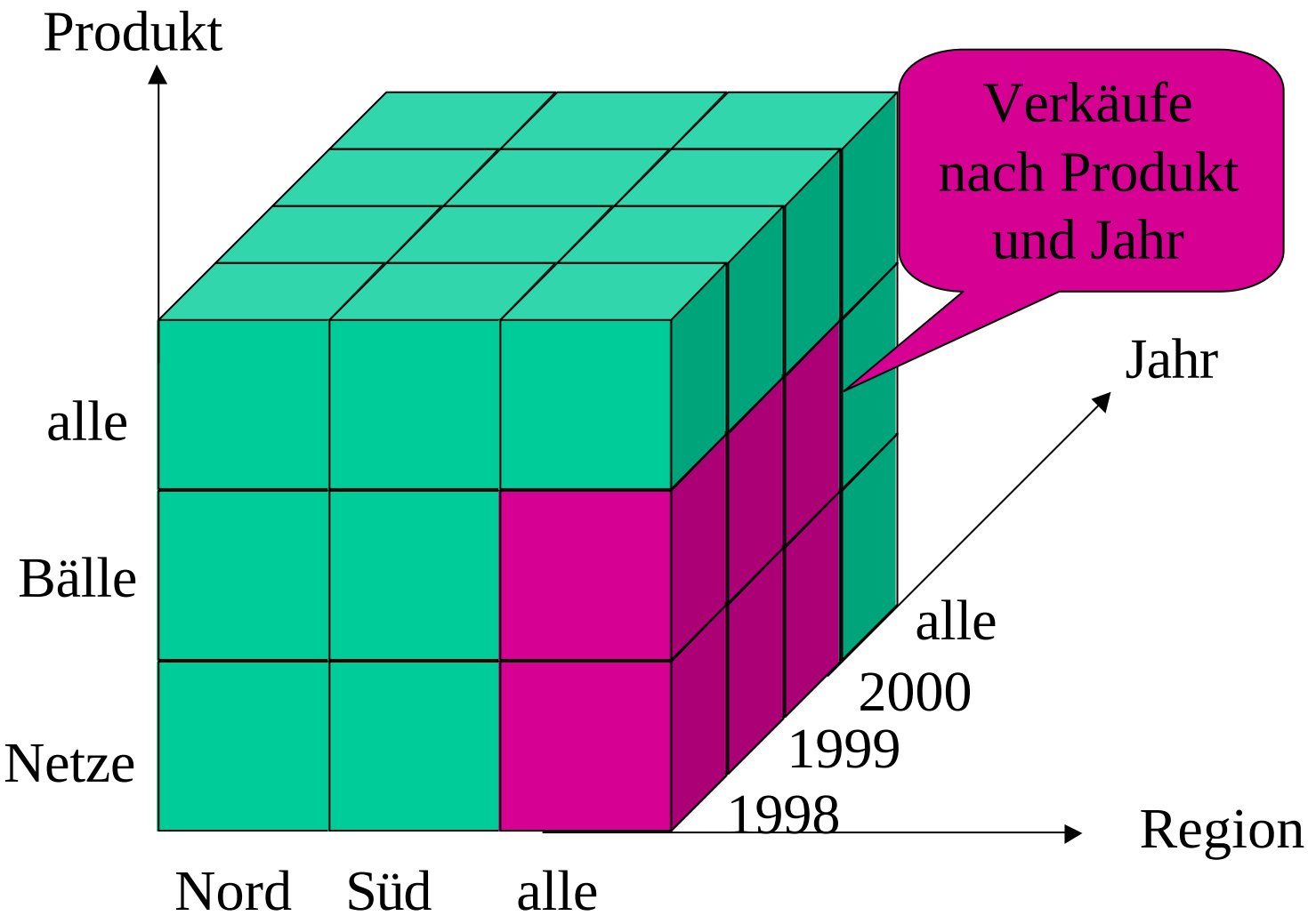
```
SELECT  f.region, z.monat, sum(a.preis * f.anzahl)
FROM    Aufträge a, Zeit z, Filiale f
WHERE  a.filiale = f.bez AND a.datum = z.datum
GROUP BY f.region AND z.monat
```

Süd	Mai	2500
Nord	Juni	1200
Süd	Oktober	5200
Nord	Oktober	600

Drill-Down und Roll-Up

- Hinzunahme eines Merkmals führt zu detaillierteren Ergebnissen (d.h. mehr Antworten)
- Wegnahme eines Merkmals führt zu gröberen, verdichteten Ergebnissen (d.h. weniger Antworten)
- GUIs erlauben „Navigation“ durch Ergebnisse
  - Drill-Down: mehr Detail
  - Roll-Up: größere Verdichtung
- Navigation auch entlang einer Hierarchie möglich: Ort statt Region erhöht den Detailgrad

Visualisierung als Würfel



Moving Sums, ROLLUP

- Manchmal möchte man entlang einer Hierarchie gruppieren und Zwischensummen bilden
- Beispiel: Man gruppiert nach Land, Region, Ort
  Aber man möchte die Kennzahlen für ein gesamtes Land und für eine Region auch haben
- Hierzu dient der sogenannte ROLLUP Operator
- Vorsicht: Die Reihenfolge ist wichtig!!!
- Hintergrund: Spreadsheets können das sehr gut
- Das Ergebnis ist aber immernoch eine Tabelle (ROLLUP funktioniert auch rein relational)

Beispiel Rollup alla IBM UDB

```
SELECT Land, Region Ort, sum(preis*anz)
FROM    Aufträge a, Filiale f
WHERE  a.filiale = f.bez
GROUP BY  ROLLUP(Land, Region, Ort)
ORDER BY Land, Region, Ort;
```

Funktioniert natürlich auch mit AVG etc. nicht nur SUM.

Ergebnis von ROLLUP

D	Nord	Köln	1000
D	Nord	(null)	1000
D	Süd	Mainz	3000
D	Süd	München	200
D	Süd	(null)	3200
D	(null)	(null)	4200

Cube Operator

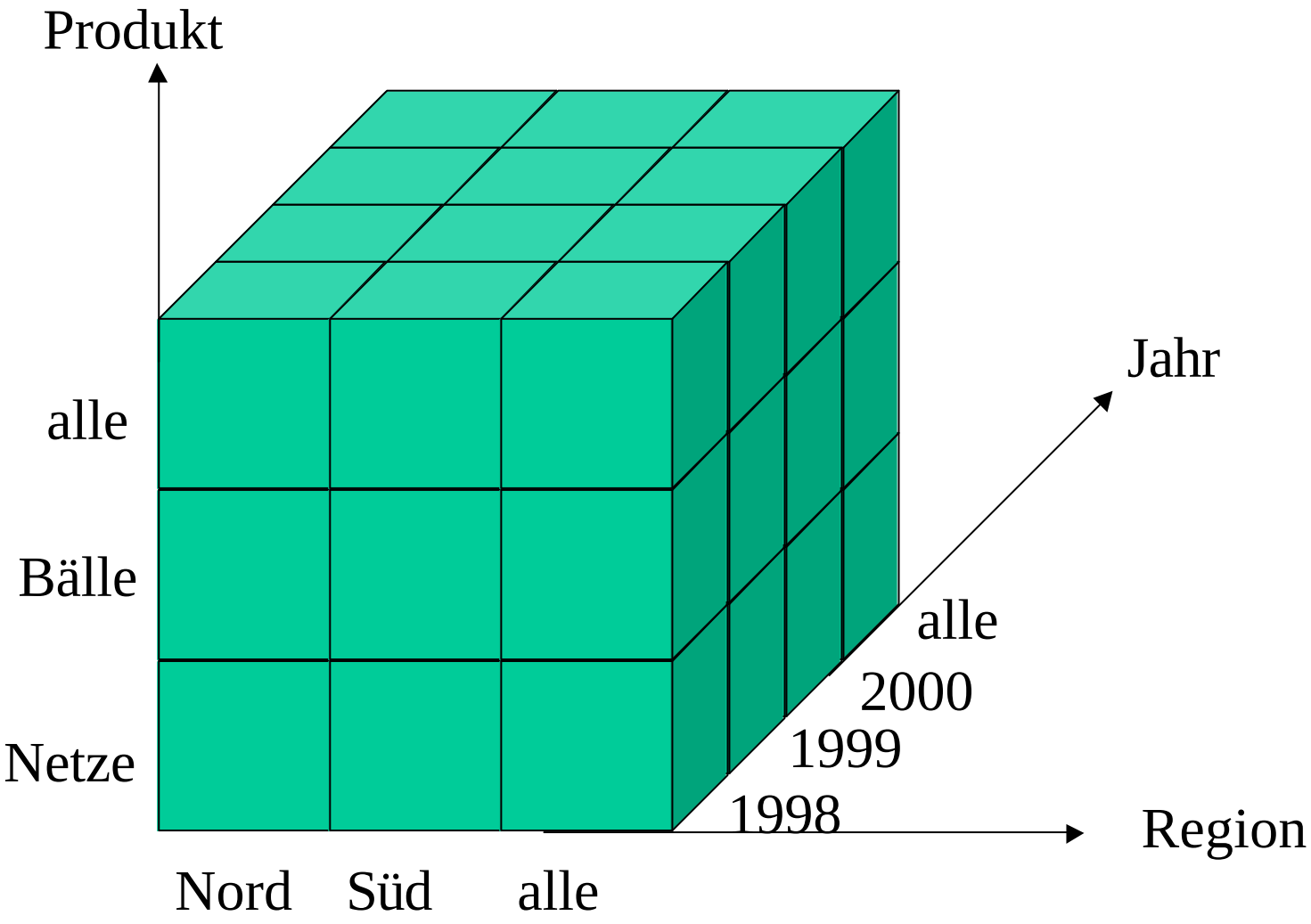
- Operator, um den kompletten Datenwürfel zu berechnen
- Ergebnis ist eine Relation mit „(null)“ Werten

```
SELECT produkt, jahr, region, sum(preis * anz)
FROM   Aufträge
GROUP BY CUBE(produkt, jahr, region);
```

Ergebnis des Cube Operators

Produkt	Region	Jahr	Umsatz
Netze	Nord	1998	...
Bälle	Nord	1998	...
(null)	Nord	1998	...
Netze	Süd	1998	...
Bälle	Süd	1998	...
(null)	Süd	1998	...
Netze	(null)	1998	...
Bälle	(null)	1998	...
(null)	(null)	1998	...

Visualisierung als Würfel



Implementierung

- ROLAP – Grundlage ein relationales System
  - Spezielle Star-Join Techniken
  - Bitmap Indexe
  - Partitionierung der Daten nach Zeit (zum Löschen)
  - Materialisierte Sichten
- MOLAP – spezielle Multidimensionale Systeme
  - Implementierung des Würfels als Array
  - Vorteil: potentiell schnell
  - Problem: Das Array ist sehr dünn besetzt
- Religionskrieg (kommerziell nur ROLAP)

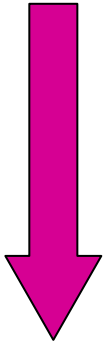
Materialisierte Sichten

- Man kann aus den Ergebnissen einer Anfrage manchmal sehr schnell andere Anfragen auswerten
- Prinzip: Subsumption
  Die Menge aller deutschen Forscher ist eine Teilmenge der Menge aller Forscher
- Besonders attraktiv für Gruppierungen

Materialisierte Sichten

```
SELECT produkt, jahr, region, sum(preis * anz)
FROM    Aufträge
GROUP BY produkt, jahr, region;
```

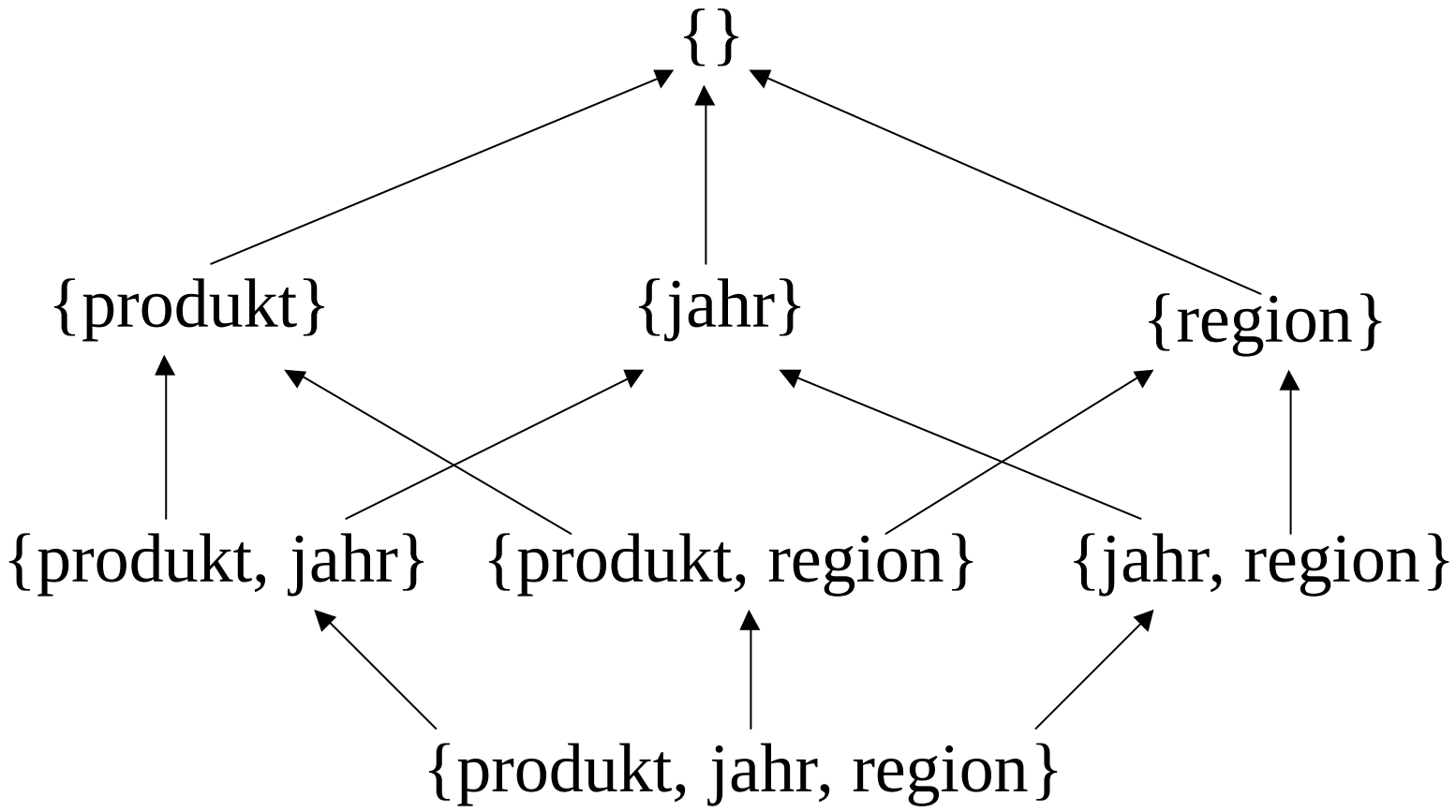
GROUP BY
product, jahr



Materialisierte Sicht

```
SELECT produkt, jahr, sum(preis * anz)
FROM    Aufträge
GROUP BY produkt, jahr;
```

Berechnungsgraph des Cube

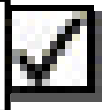


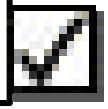
Online Aggregation

- Schnell Abschätzungen bekommen
- Lösung verbessert sich mit der Laufzeit
- Basiert auf (unabhängige) Stichproben
- (Noch) keine kommerzielle Verbreitung

```
SELECT  kunde, avg(preis)
FROM    Aufträge
GROUP BY kunde
```

Kunde	Avg	+/-	Conf
Heinz	1375	5%	90%
Ute	2000	5%	90%
Karin	-	-	-





Ranking, Top N

- Gib mir die 10 Produkte, die sich im Jahr 2000 am besten verkauft haben
- (Syntax: Carey, Kossmann 1997)

```
SELECT z.product, sum(z.preis * z.anz) as umsatz
FROM    Aufträge a, Zeit z
WHERE a.datum = z.datum AND z.jahr = 2000
ORDER BY umsatz DESC
STOP AFTER 10;
```

Top N in Datenbankprodukten

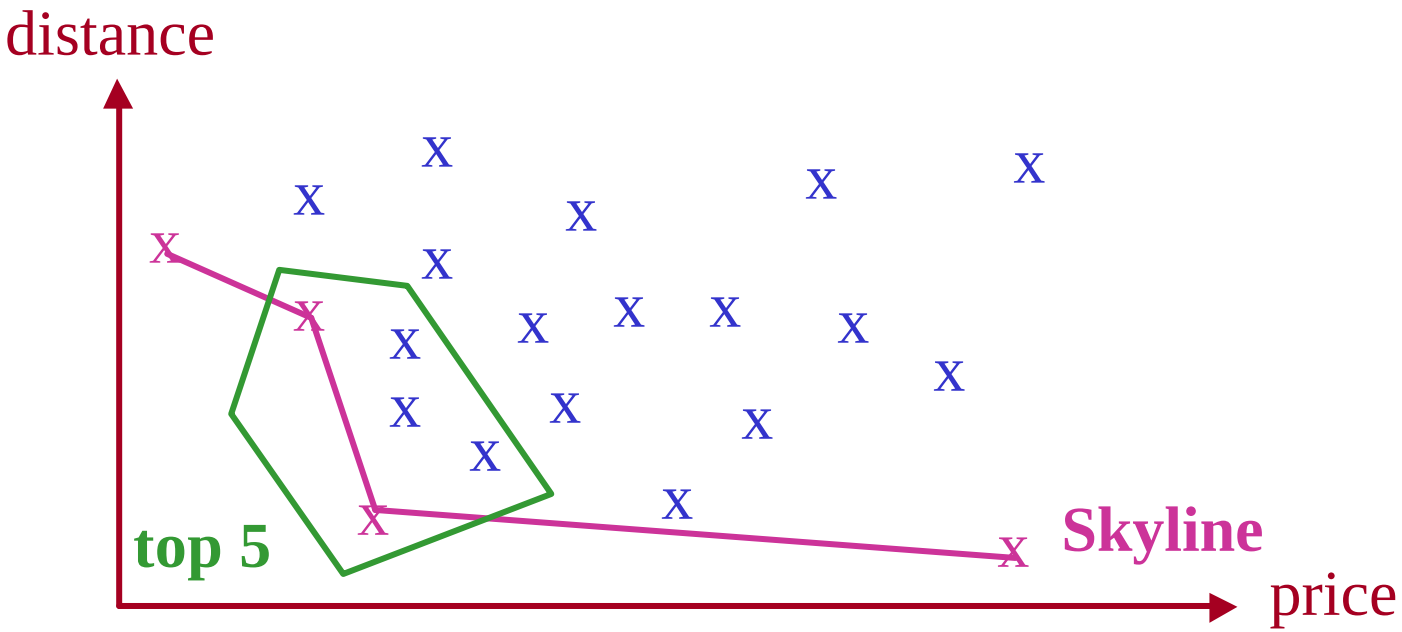
- Microsoft SQL Server
  - Separate **set rowcount N** Klausel
- Oracle
  - **rownum < N** Prädikat in WHERE Klausel
- IBM UDB
  - **fetch first N rows only**
- Fragen:
  - Kann man den Rank auch ausgeben?
  - Was passiert, wenn mehrere Produkte den gleichen Umsatz erzeugen?

Skyline Queries

- What happens if the user is interested in hotels which are cheap and near the lake
- that is, there is more than one target
- 1. Try: top N query with scoring function
  - e.g., ORDER BY price * *x* + distance * *y*
  - how do you set *x*, *y*?
  - result contains many similar hotels
- 2. Try: Skyline query

Skyline Queries

- Return all incomparable hotels



Skyline

- Suche billige Hotels nahe am Strand
- (Syntax: Börszöny, Kossmann, Stocker 2001)

```
SELECT  *
FROM    Hotels h
WHERE  h.island = „Bahamas“
SKYLINE OF price MIN, distance MIN;
```

Skyline of New York

Suche Gebäude, die Hoch sind und nahe am Fluss
und zwar für jeden Straßenzug separat.

```
SELECT *
FROM    Buildings
WHERE city = „New York“
SKYLINE OF heigth MAX, distance MIN, street DIFF;
```

Data Mining

- Statische Auswertungen auf dem Datenbestand
- Ziel: Antworten zu bekommen, ohne dass man die Fragen stellen muss
- Verschiedene Arten und Anwendungen
  - Association Rule Mining
  - Clustering
  - Outlyer (Fraud) Detection
- Eigener Markt für besondere Tools
  z.B. IBM Intelligent Miner

Association Rule Mining
(Warenkorbanalyse)

Wenn jemand Windeln kauft, dann kauft er auch Bier

- Confidence: In wieviel Prozent der Warenkörbe mit Bier liegen auch Windeln (typisch sind 50%)
- Support: In wieviel Prozent der Warenkörbe liegen Bier und Windeln (typisch sind 1%)
- Association Rule Mining: Finde alle Regeln bei gegebenem (minimalem) Confidence und Support
- Bedeutung: Angeblich hat durch diese Art von Analyse WalMart Bierdisplays bei den Windelregalen eingerichtet

Clustering

- Ziel: Klassifikation von Objekten
- Variante 1: Es gibt ein Klassifikationsschema
  - Aufgabe: neue Objekte zu klassifizieren
- Variante 2: Es gibt kein Klassifikationsschema
  - Aufgabe: Klassifikationsschema erstellen
- Anwendungen: Versicherungen, Genomentschlüsselung, Aktienanalyse, ...
- Grundlegende Operation: Nearest Neighbor Search

Outlier Detection

- Finde Objekte, die nicht in das Klassifikationsschema passen
- Oder finde Objekte, die von der Norm abweichen
- Anwendungen: Erkennen von Attacken auf ein System (Analyse des Logs), Missbrauch von Telefonen (z.B. nach Handydiebstahl)
- Noch relativ offenes Forschungsgebiet