Datenbanksysteme 2009

Kapitel 17:
Data Warehouse

Oliver Vornberger

Institut für Informatik
Universität Osnabrück

OLTP versus OLAP

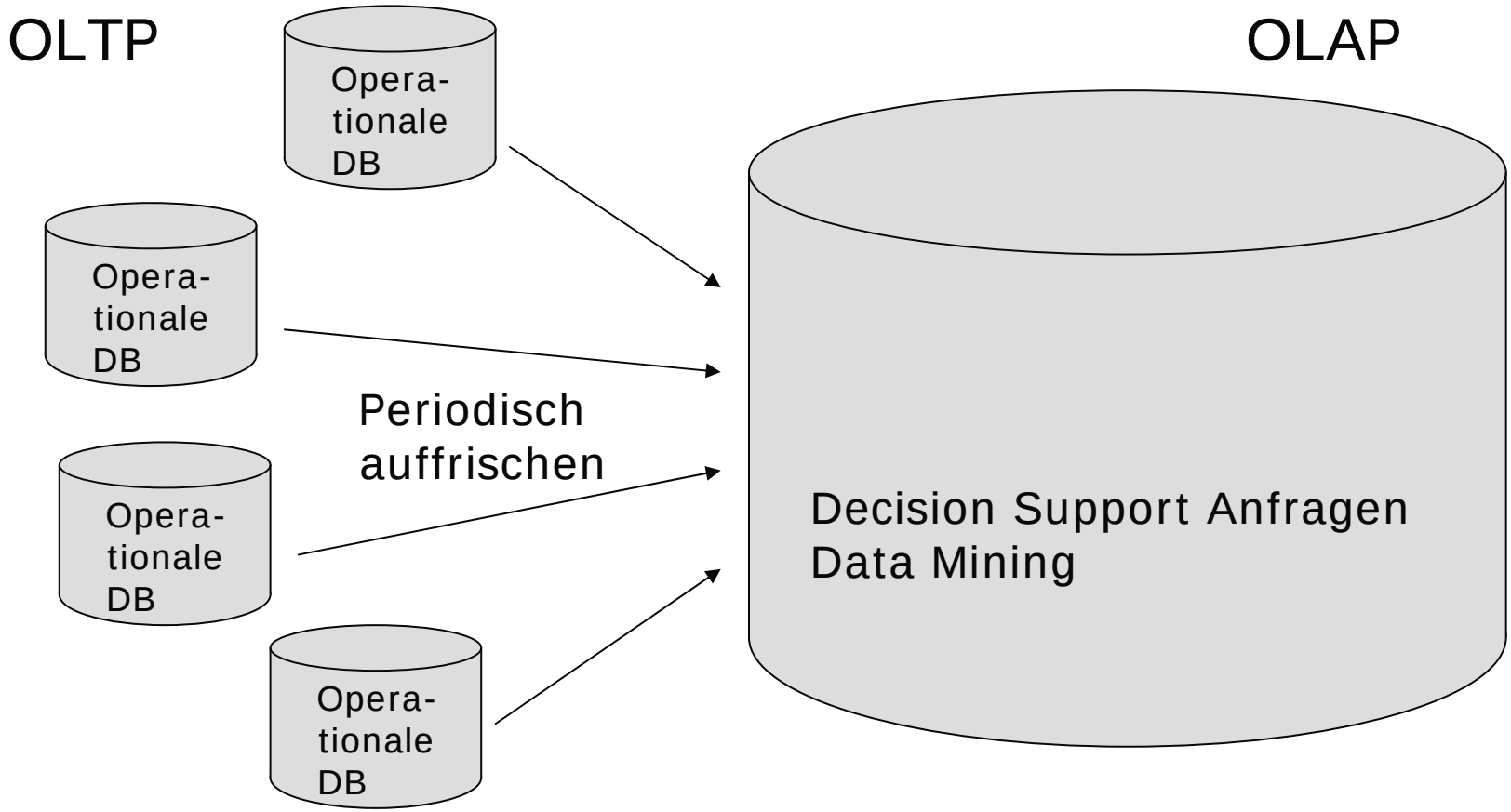
OLTP (Online Transaction Processing)

- z.B. Flugreservierung, Handelsunternehmen
- kleine, kurze Transaktionen
- jeweils auf jüngstem Zustand

OLAP (Online Analytical Processing)

- z.B. Auslastung der Transatlantikflüge
- z.B. Umsatzrückgang im 3. Quartal
- große Datenmengen
- historische Daten
- Grundlage für Decision Support Systeme

Zusammenspiel OLTP/OLAP



Sternschema

Handelsunternehmen

Krankenkasse

Aus prä gung

Verkäufe					
VerkDatum	Filiale	Produkt	Anzahl	Kunde	Verkäufer
30-Jul-96	Passau	1347	1	4711	825
...	...	...	...	...	...

Filialen			
Filialenkennung	Land	Bezirk	...
Passau	D	Bayern	...
...	...	...	...

Kunden			
KundenNr	Name	wiealt	...
4711	Kemper	38	...
...	...	...	...

Verkäufer					
VerkäuferNr	Name	Fachgebiet	Manager	wiealt	...
825	Handyman	Elektronik	119	23	...
...	...	...	...	...	...

Zeit								
Datum	Tag	Monat	Jahr	Quartal	KW	Wochentag	Saison	...
...	...	...	...	...	...	...	...	...
30-Jul-96	30	Juli	1996	3	31	Dienstag	Hochsommer	...
...	...	...	...	...	...	...	...	...
23-Dec-97	27	Dezember	1997	4	52	Dienstag	Weihnachten	...
...	...	...	...	...	...	...	...	...

Produkte					
ProduktNr	Produkttyp	Produktgruppe	Produkthauptgruppe	Hersteller	...
1347	Handy	Mobiltelekom	Telekom	Siemens	...
...	...	...	...	...	...

Star Join

Welche Handys (d.h. von welchen Herstellern) haben junge Kunden in den bayrischen Filialen zu Weihnachten 1996 gekauft ?

```
select p.Hersteller, sum(v.Anzahl) as Anzahl
from Verkäufe v, Filialen f, Produkte p, Zeit z, Kunden k
where z.Saison = 'Weihnachten'
and z.Jahr       = 1996
and k.wiealt     < 30
and p.Produkttyp = 'Handy'
and f.Bezirk     = 'Bayern'
and v.VerkDatum  = z.Datum
and v.Produkt    = p.ProduktNr
and v.Filiale    = f.Filialenkennung
and v.Kunde      = k.KundenNr
group by Hersteller;
```

Verdichtungsgrad

Drill down:

- mehr Attribute in group-by-Klausel
- geringere Verdichtung

Roll up:

- weniger Attribute in group-by-Klausel
- stärkere Verdichtung

Handyverkäufe nach Herstellern

```
select p.Hersteller, sum(v.Anzahl) as Anzahl
from   Verkäufe v, Produkte p
where  v.Produkt = p.ProduktNr
and    p.Produkttyp = 'Handy'
group  by p.Hersteller;
```

Drill down längs Zeitachse

```
select p.Hersteller, z.Jahr,
        sum(v.Anzahl) as Anzahl
from   Verkäufe v, Produkte p
where  v.Produkt = p.ProduktNr
and    p.Produkttyp = 'Handy'
group  by p.Hersteller, z.Jahr;
```

Roll up längs Hersteller

```
select      z.Jahr, sum(v.Anzahl) as Anzahl
from        Verkäufe v, Produkte p
where       v.Produkt = p.ProduktNr
and         p.Produkttyp = 'Handy'
group by    z.Jahr;
```

Analyse nach Dimensionen

Handyverkäufe nach Hersteller und Jahr		
Hersteller	Jahr	Anzahl
Siemens	1994	2.000
Siemens	1995	3.000
Siemens	1996	3.500
Motorola	1994	1.000
Motorola	1995	1.000
Motorola	1996	1.500
Bosch	1994	500
Bosch	1995	1.000
Bosch	1996	1.500
Nokia	1994	1.000
Nokia	1995	1.500
Nokia	1996	2.000

Drilldown längs Jahr

Handyverkäufe nach Jahr	
Jahr	Anzahl
1994	4.500
1995	6.500
1996	8.500

Handyverkäufe nach Hersteller	
Hersteller	Anzahl
Siemens	8.500
Motorola	3.500
Bosch	3.000
Nokia	4.500

Roll-Up längs Hersteller / Jahr

Ultimative Verdichtung

```
select      sum(v.Anzahl) as Anzahl
from        Verkäufe v, Produkte p
where       v.Produkt = p.ProduktNr
and         p.Produkttyp = 'Handy'
```

Handyverkäufe
Anzahl
19.500

Materialisierung

gewünscht:
Handy-Verkäufe nach Jahr und Hersteller

Hersteller \ Jahr	1994	1995	1996	Σ
Siemens	2.000	3.000	3.500	8.500
Motorola	1.000	1.000	1.500	3.500
Bosch	500	1.000	1.500	3.000
Nokia	1.000	1.500	2.000	4.500
Σ	4.500	6.500	8.500	19.500

Handy2Cube

Wunsch:

verschieden aggregierte Teilergebnise vorberechnen:

```
create table Handy2DCube(
      Hersteller  varchar(20),
      Jahr        integer,
      Anzahl      integer
);
```

Berechnung Hersteller + Jahr

```
insert into  Handy2DCube

(select p.Hersteller, z.Jahr, sum(v.Anzahl)
from  Verkäufe v, Produkte p, Zeit z
where v.Produkt = p.ProduktNr
and p.Produkttyp = 'Handy'
and   v.VerkDatum = z.Datum
group by z.Jahr, p.Hersteller)


union (select null, z.Jahr, sum(v.Anzahl)
from  Verkäufe v, Produkte p, Zeit z
where v.Produkt = p.ProduktNr
and p.Produkttyp = 'Handy'
and   v.VerkDatum = z.Datum
group by z.Jahr)

union (select p.Hersteller, null, sum(v.Anzahl)
from  Verkäufe v, Produkte p
where v.Produkt = p.ProduktNr
and p.Produkttyp = 'Handy'
group by p.Hersteller)

union (select null, null, sum(v.Anzahl)
from  Verkäufe v, Produkte p
where v.Produkt = p.ProduktNr
and p.Produkttyp = 'Handy');
```

Handy2DCube		
Hersteller	Jahr	Anzahl
Siemens	1994	2.000
Siemens	1995	3.000
Siemens	1996	3.500
Motorola	1994	1.000
Motorola	1995	1.000
Motorola	1996	1.500
Bosch	1994	500
Bosch	1995	1.000
Bosch	1996	1.500
Nokia	1994	1.000
Nokia	1995	1.500
Nokia	1996	2.000
null	1994	4.500
null	1995	6.500
null	1996	8.500
Siemens	**null**	8.500
Motorola	**null**	3.500
Bosch	**null**	3.000
Nokai	**null**	4.500
null	**null**	19.500

Cube-Operator

```
select p.Hersteller, z.Jahr, f.Land, sum(Anzahl)
                                          as Anzahl
from   Verkäufe v, Produkte p, Zeit z, Filialen f
where  v.Produkt    = p.ProduktNr
and    p.Produkttpy = 'Handy'
and    v.VerkDatum  = z.Datum
and    v.Filiale    = f.Filialenkennung
group  by z.Jahr, p.Hersteller, f.Land
with cube;
```

Erzeugt bei k Attributen in group-by-Klausel
2^k Teilmengen (inkl. Null-Werte) und vereinigt sie.

Ergebnis des Cube Operators

Handy3DCube			
Hersteller	Jahr	Land	Anzahl
Siemens	1994	D	800
Siemens	1994	A	600
Siemens	1994	CH	600
Siemens	1995	D	1.200
Siemens	1995	A	800
Siemens	1995	CH	1.000
Siemens	1996	D	1.400
...	...	...	...
Motorola	1994	D	400
Motorola	1994	A	300
Motorola	1994	CH	300
...	...	...	...
Bosch	...	...	...
...	...	...	...
null	1994	D	...
null	1995	D	...
...	...	...	...
Siemens	**null**	**null**	8.500
...	...	...	...
null	**null**	**null**	19.500

MySQL: WITH ROLLUP

```
select Hersteller, Jahr, sum(Umsatz) as Umsatz
from Handys
group by Hersteller, Jahr
with Rollup
```

Hersteller	Jahr	Umsatz
Bosch	1994	500
Bosch	1995	1000
Bosch	1996	1500
Bosch	*(null)*	3000
Motorola	1994	1000
Motorola	1995	1000
Motorola	1996	1500
Motorola	*(null)*	3500
Nokia	1994	1000
Nokia	1995	1500
Nokia	1996	2000
Nokia	*(null)*	4500
Siemens	1994	2000
Siemens	1995	3000
Siemens	1996	3500
Siemens	*(null)*	8500
(null)	*(null)*	19500

Aqua Data Studio

Data Mining

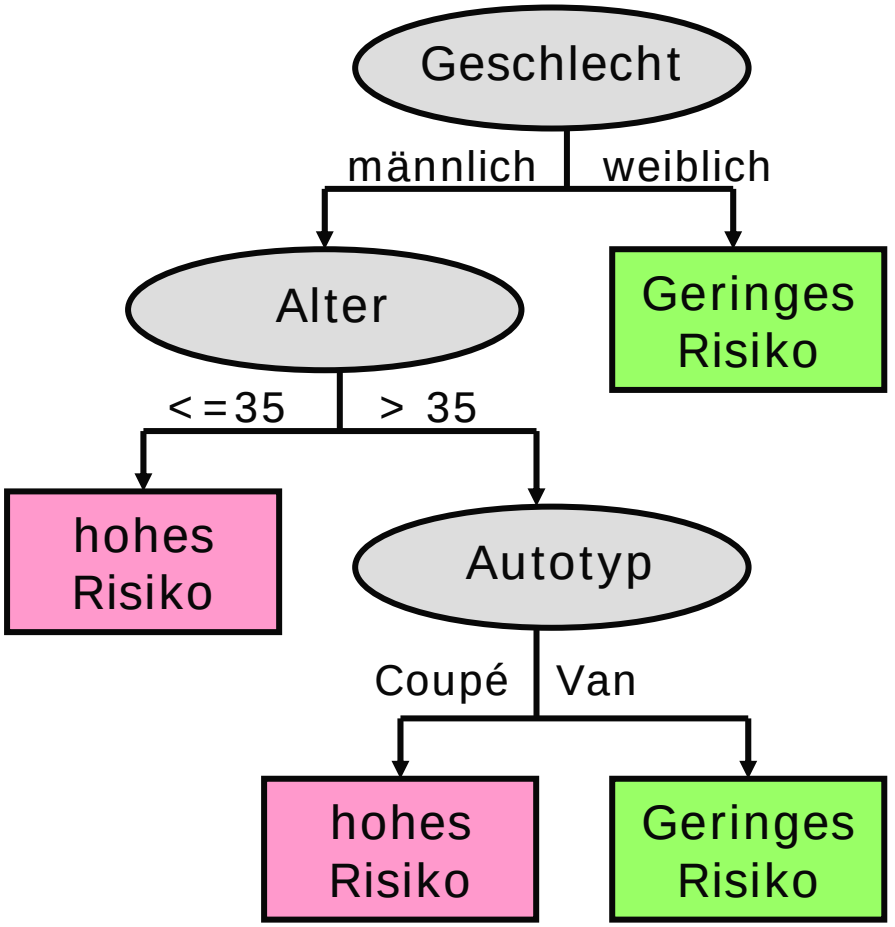
Datenbestand durchsuchen nach (bisher unbekannten) Zusammenhängen (Knowledge Discovery)

- Klassifikation von Objekten:
  z.B. Risikoabschätzung bei Versicherungspolicen
  männlich + alt + Coupé $\Rightarrow$ hohes Risiko

- Assoziationsregeln:
  z.B. Analyse des Kaufverhaltens
  "Wer PC kauft, kauft auch Drucker"

Beispiel für Klassifikation

Alter	Geschlecht	Autotyp	Schaden
45	**w**	**Van**	**gering**
18	**w**	**Coupé**	**gering**
22	**w**	**Van**	**gering**
19	**m**	**Coupé**	**hoch**
38	**w**	**Coupé**	**gering**
24	**m**	**Van**	**Gering**
40	**m**	**Coupé**	**hoch**
40	**m**	**Van**	**gering**
…	…	…	…



Beispiel für Assoziationsregeln

- Suche in Warenkörben nach gängigen Kombinationen.
- Leite daraus Schlussfolgerungen ab

Frequent Itemset in ≥ k Einkäufen

TransID	Produkt
111	Drucker
111	Papier
111	PC
111	Toner
222	PC
222	Scanner
333	Drucker
333	Papier
333	Toner
444	Drucker
444	PC
555	Drucker
555	Papier
555	PC
555	Scanner
555	Toner

Frequent Itemset-Kandidat	Anzahl
{Drucker}	4
{Papier}	3
{PC}	4
{Scanner}	2
{Toner}	3
{Drucker, Papier}	3
{Drucker, PC}	3
{Drucker, Scanner}	
{Drucker, Toner}	3
{Papier, PC}	2
{Papier, Scanner}	
{Papier, Toner}	3
{PC, Scanner}	
{PC, Toner}	2
{Scanner, Toner}	
{Drucker, Papier, PC}	
{Drucker, Papier, Toner}	3
{Drucker, PC, Toner}	
{Papier, PC, Toner}	

Berechnung der Confidence

Sei F ein frequent itemset

support(F) := Anzahl des Vorkommens / Gesamtzahl

Betrachte alle disjunkten Zerlegungen in L und R

Die Regel $L \Rightarrow R$ hat

$$confidence\ (L \Rightarrow R) = support(F) / support(L)$$

Beispiel: {Drucker} $\Rightarrow$ {Papier, Toner}

$$confidence = \frac{support(\{\text{Drucker, Papier, Toner}\})}{support(\{\text{Drucker}\})} = \frac{3/5}{4/5} = 0.75$$

75 % der Kunden, die einen Drucker gekauft haben, haben auch Papier und Toner gekauft