

**Übung zur Vorlesung
„Einsatz und Realisierung von Datenbanksystemen“
im Sommersemester 2007**

Richard Kuntschke (richard.kuntschke@in.tum.de)

Lösungen zu Blatt 7

Aufgabe 1

Die folgenden Überlegungen geben eine ungefähre Abschätzung der Größe eines repräsentativen Handelsunternehmens wieder und stellen keine offiziellen Daten der Unternehmen dar.

Wir wollen diese Abschätzung am Beispiel von Amazon.com durchführen. In den Jahren 2002-2004 erzielte Amazon einen Umsatz von insgesamt rund 16,1 Mrd. \$. Wir gehen von den weiteren (hypothetischen) Rahmenbedingungen aus:

1. Der Preis für einen verkauften Artikel beträgt im Durchschnitt 10 \$.
2. Ein Kunde kauft pro Jahr im Schnitt fünf Artikel bei Amazon.com ein.
3. Amazon.com bietet 100.000 unterschiedliche Produkte an.

Ausgehend von diesen Rahmenbedingungen können wir die Größe der einzelnen Tabellen (vgl. Abbildung 1) abschätzen:

Tabelle *Verkäufe*: Ein Tupel der Faktentabelle referenziert über Fremdschlüsselbeziehungen Tupel in den fünf Dimensionstabellen. Für jede Referenz (Filiale wird nicht als **varchar** gespeichert) werden 4 Byte benötigt. Das Attribut *Anzahl* belegt ebenfalls 4 Byte.

Verkäufe hat etwa $16,1 \text{ Mrd.} / 10 = 1,61 \text{ Mrd.}$ Einträge. Daraus ergibt sich eine Größe von $1,61 \text{ Mrd.} \cdot 6 \cdot 4 \text{ Byte} \approx 36 \text{ GB}$.

Tabelle *Filialen*: Da Amazon.com ein Online-Händler ist (Filialen sind hier die Repräsentationen in den jeweiligen Ländern, also Amazon.de, Amazon.fr etc.), ist die Ausprägung von *Filialen* verhältnismäßig klein und wird im Folgenden vernachlässigt.

Tabelle *Kunden*: Anders als in Abbildung 1 nehmen wir an, ein Kundeneintrag setzt sich zusammen aus einer Kundennummer (4 Byte), einem Namen (30 Byte), einem Geburtsdatum (4 Byte) und einer Adresse (70 Byte). Nach unserer Schätzung gibt es $1,61 \text{ Mrd.} / (3 \cdot 5) = 107 \text{ Mio.}$ Kunden. Damit ist die Größe der Ausprägung ungefähr $107.000.000 \cdot 108 \text{ Byte} \approx 10,8 \text{ GB}$.

Tabelle *Verkäufer*: Produkte können zwar über Amazon nicht nur gekauft, sondern auch verkauft werden. Die Anzahl der Verkäufer wird aber (im Vergleich zur Anzahl der Kunden) als vernachlässigbar klein angenommen.

Verkäufe					
VerkDatum	Filiale	Produkt	Anzahl	Kunde	Verkäufer
25-Jul-00	Passau	1347	1	4711	825
...

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

Kunden			
KundenNr	Name	wiealt	...
4711	Kemper	43	...
...

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

Zeit								
Datum	Tag	Monat	Jahr	Quartal	KW	Wochentag	Saison	...
...
25-Jul-00	25	Juli	2000	3	30	Dienstag	Hochsommer	...
...
18-Dec-01	18	Dezember	2001	4	52	Dienstag	Weihnachten	...
...

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

Abbildung 1: Relationen des Sternschemas für ein Handelsunternehmen

Tabelle *Zeit*: Die in Abbildung 1 dargestellten acht Einträge der Relation *Zeit* werden über Identifikatoren (z.B. „2“ statt „Dienstag“) realisiert. Jeder Identifikator benötigt 4 Byte. Damit ergibt sich für drei Jahre (wenn jeder Tag einen Eintrag darstellt) eine ungefähre Größe von

$$3 \cdot 365 \cdot 8 \cdot 4 \text{ Byte} \approx 34 \text{ kB.}$$

Tabelle *Produkte*: Wie angegeben, gehen wir von ca. 100.000 Produkten aus. Für eine Produktbeschreibung inklusive eines kleinen Fotos nehmen wir an, dass im Durchschnitt 8 kB benötigt werden. Damit ergibt sich eine Größe von

$$100.000 \cdot 8 \text{ kB} \approx 0,76 \text{ GB.}$$

Nach unserer Rechnung kommen wir daher auf eine geschätzte Gesamtgröße von

$$36 \text{ GB} + 10,8 \text{ GB} + 34 \text{ kB} + 0,76 \text{ GB} \approx 47,6 \text{ GB.}$$

für das Data Warehouse von Amazon.com.

Diese Abschätzung repräsentiert aber sicherlich nur einen verhältnismäßig kleinen Ausschnitt des Datawarehouses eines Handelsunternehmens wie Amazon.com. Weitere Data Mining-Anwendungen

stellen etwa Recommender-Systeme dar, die das Kundenverhalten auswerten und Kaufempfehlungen erstellen, wozu Benutzeraktionen (im Wesentlichen jeder Klick) vom System mitprotokolliert werden.

Aufgabe 2

Ein Eintrag im Datenwürfel repräsentiert die Stückzahl eines in einem bestimmten Monat, von einem Kunden gekauften Produkts. Ein Eintrag ist vom Datentyp *Integer* und benötigt daher 4 Byte.

Wie in Aufgabe 1 motiviert, gehen wir von folgenden Eckdaten aus:

Dimension *Kunden*: $107 \cdot 10^6$
Dimension *Produkte*: 10^5
Dimension *Monate*: $12 \cdot 3 = 36$
Bytes pro Eintrag: 4

Damit hat der Datenwürfel folgende Größe:

$$107 \cdot 10^6 \cdot 10^5 \cdot 36 \cdot 4 \text{ Byte} \approx 1541 \cdot 10^{12} \text{ Byte} \approx 1401 \text{ TB} \approx 1,4 \text{ PB.}$$

Dieser Wert schätzt jedoch nur den Speicherbedarf für die Einträge des Datenwürfels ab. Würde man den Würfel relational abspeichern, müsste man zusätzlich noch den Platzbedarf für die Dimensionsattribute eines jeden Tupels berücksichtigen.

Aufgrund des immensen Speicherbedarfs ist davon auszugehen, dass für diese spezielle Anfrage das Ergebnis nicht materialisiert wird. Zudem sind die meisten Einträge im Datenwürfel leer. Bei einer relationalen Darstellung des Würfels würde man diese Einträge weglassen, so dass die obige Abschätzung nur dann gilt, wenn man jeden Eintrag (also auch leere Felder) abspeichert.

Aufgabe 3

Die SQL-Anfrage zur Berechnung des Datenwürfels sieht wie folgt aus:

```
select p.Produkttyp, f.Bezirk, k.wiealt, sum(v.Anzahl)
from Verkäufe v, Produkte p, Filialen f, Kunden k, Zeit z
where v.Produkt = p.ProduktNr and v.Kunde = k.KundenNr
      and v.Filiale = f.Filialenkennung and v.VerkDatum = z.Datum
      and z.Saison = 'Hochsommer' and f.Land = 'D'
group by cube(p.Produkttyp, f.Bezirk, k.wiealt);
```

Abbildung 2 zeigt den Aufbau des Datenwürfels schematisch. In einem inneren Datenpunkt des Würfels steht die Summe der Verkäufe je Produkttyp, Bezirk und Alter der Käufergruppe. Die mit dem Summenzeichen Σ dargestellten Datensätze illustrieren die Aggregationen bzgl. der jeweiligen Dimensionen.

Aufgabe 4

Wir gehen von einer konsistenten Datenbasis aus, die nur gültige Wahlzettel berücksichtigt, also insbesondere solche, die eine gültige Erst- und Zweitstimme enthalten. Damit kann zur Berechnung der Wahlbeteiligung je Wahllokal entweder nur die Erst- oder nur die Zweitstimme betrachtet werden:

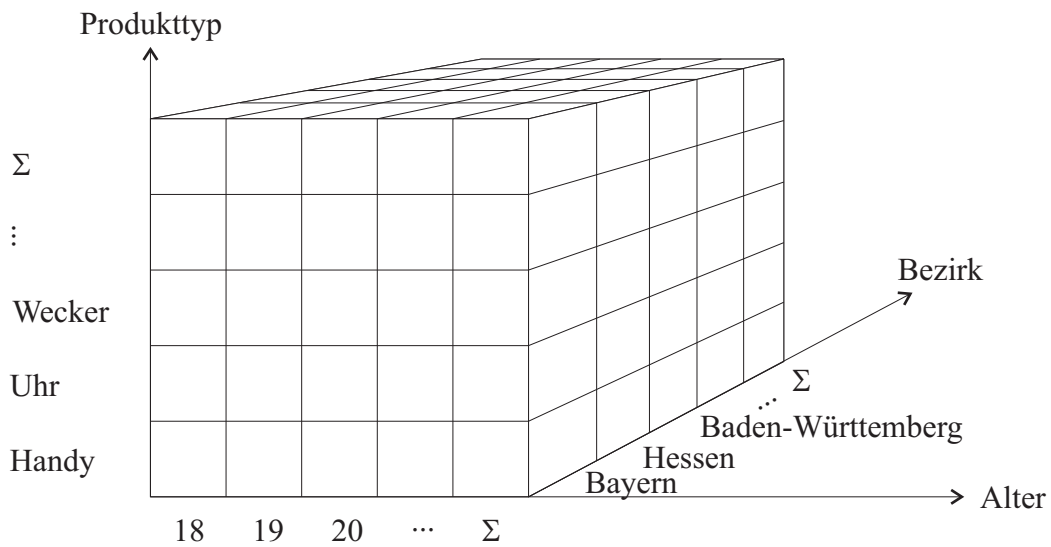


Abbildung 2: Würfeldarstellung der Verkaufszahlen nach Alter der Käufer, Produkttyp und Bezirk

```
create view Wähler2005 as
select wb.Nr, wb.Wahlberechtigte, sum(e.Stimmen) as Wähler
from Erststimmen e, Wahlbezirke wb
where e.Jahr = 2005 and e.Wahlbezirk = wb.Nr
group by wb.Nr, wb.Wahlberechtigte;
```

Aufbauend auf dieser View kann nun die Wahlbeteiligung in entsprechender Detailtiefe ermittelt werden:

```
select b.Name, wk.Nr, wb.Nr,
       sum(w05.Wähler), sum(w05.Wahlberechtigte),
       cast(sum(w05.Wähler) as float)/
       cast(sum(w05.Wahlberechtigte) as float) as Beteiligung
from Bundesländer b, Wahlkreise wk, Wahlbezirke wb, Wähler2005 w05
where b.Name = wk.Bundesland
       and wk.Nr = wb.Wahlkreis and wb.Nr = w05.Nr
group by cube(b.Name, wk.Nr, wb.Nr);
```