

Datenbanken
Wintersemester 09/10
 Prof. Dr. W. May

4. Übungsblatt: Physischer Entwurf, Interna, Optimierung

Besprechung voraussichtlich am 26.1.2010

Aufgabe 1 (Join-Algorithmen) Gegeben sind zwei Tabellenschemata R und S :

R		
<u>A</u>	<u>B</u>	C
VARCHAR2(3)	NUMBER	VARCHAR2(20)

S		
<u>E</u>	<u>F</u>	<u>G</u>
VARCHAR2(3)	VARCHAR2(25)	VARCHAR2(25)

Die Aufgabe behandelt das Join $R \bowtie_{R.A=S.E} S$.

Physikalisches Datenbankschema:

- Schlüssel von R ist (A, B) , Schlüssel von S ist (E, F) .
- Der Wertebereich von $R.A$ ist gleich dem Wertebereich von $S.E$, d.h. $\pi[A](R) = \pi[E](S)$.
- Die Tupel von R sind in beliebiger Reihenfolge gespeichert.
- Die Tupel von S sind so gespeichert, dass Tupel mit demselben E -Wert benachbart gespeichert sind. Eine weitergehende Ordnung existiert nicht.

Daten:

- R enthält 2.000.000 Tupel [Hinweis: diese Schreibweise bedeutet in Deutschland "2 Millionen"].
- S enthält 4.000.000 Tupel.
- Die Tabellen sind wie üblich seitenweise abgelegt. Eine Seite (4KB) von R enthält 100 Tupel, eine Seite von S enthält 60 Tupel.
- Zu jedem Wert a von A gibt es genau 2 Tupel $\mu \in R$, so dass $\mu[A] = a$ ist.
- Zu jedem Wert e von E gibt es durchschnittlich 4 Tupel $\mu \in S$, so dass $\mu[E] = e$ ist.
- Im Hauptspeicher können 1000 Seiten gleichzeitig gehalten werden.

Die Anfrage $\pi[A, B, F](R \bowtie_{R.A=S.E} S)$ soll berechnet und das Ergebnis ausgegeben werden.

A. Join unter Verwendung eines Indexes

- Nehmen Sie an, dass für das Attribut $S.E$ ein B*-Baumindex angelegt ist, dessen Einträge jeweils auf das erste Tupel mit dem entsprechenden Wert von $S.E$ zeigen.
- Für den B*-Baum über $S.E$ gilt folgendes: Jedes Blatt enthält durchschnittlich 100 Einträge. Jeder innere Knoten enthält 200 Einträge.

Fragen:

a) Wieviele Werte enthält $\pi[A](R)$?

- b) Wieviele Blätter enthält der B*-Baum über $S.E$?
- c) Wie viele Ebenen hat der Baum?
Wieviele Knoten enthält der Baum insgesamt?
- d) Die Anfrage $\pi[A, B, F](R \bowtie_{R.A=S.E} S)$ soll berechnet und am Bildschirm ausgegeben werden.
Wieviele Ergebnistupel bekommt man?
- e) Beschreiben Sie, wie die Auswertung (unter Verwendung des Indexes) am besten vorgeht.
Wieviele Hintergrundspeicherzugriffe auf die Datenbank werden benötigt?
(beschreiben Sie am besten zuerst grob, wie Sie vorgehen, und analysieren Sie dann die Zugriffe)
(insgesamt 10P)

B. Hash-Join Untersuchen Sie die Anwendung eines Hash-Joins über $R.A$ bzw. $S.E$ um $\pi[A, B, F](R \bowtie_{R.A=S.E} S)$ zu berechnen.

- a) Untersuchen Sie zuerst den naiven Ansatz, dies zu tun.
- b) Optimieren Sie die Anfrage zuerst algebraisch und überlegen Sie sich dann eine naheliegende Optimierung.
- c) Kann man das Vorgehen noch weiter optimieren?

Aufgabe 2 (Optimierung, Aufwandsabschätzung (SQL)) Gegeben sei folgendes Datenbankschema (Auszug aus Mondial)

```
Country(Name, Code, Capital, Province, Area, Population)
Organization(Name, Abbreviation, Established)
Is_member(Organization, Country, Type)
```

Formulieren Sie die folgenden Anfragen in SQL:

- a) Welche Länder sind Mitglied in mehr als 60 Organisationen?
- b) Welche Länder mit einer Fläche von mehr als 500000 km² sind Mitglied in mehr als 60 Organisationen?

Untersuchen Sie bei der 2. Teilaufgabe den Aufwand (Anzahl Hintergrundspeicherzugriffe, Vergleiche zur Ausführung des Join, Vergleiche zum Test der Fläche). An welchen Stellen wäre ein (Baum- oder anderer) Index nützlich? Untersuchen Sie auch für diesen Fall den Aufwand.

Können Sie Ihre Anfrage optimieren, um damit den Aufwand weiter zu reduzieren?

Aufgabe 3 (Mengenoperationen) Seien $R(\bar{X}), S(\bar{X})$ Schemata und $\bar{Y} \subseteq \bar{X}$. Zeigen Sie oder widerlegen Sie:

- a) $\pi[\bar{Y}](R \cup S) = \pi[\bar{Y}](R) \cup \pi[\bar{Y}](S)$
- b) $\pi[\bar{Y}](R \cap S) = \pi[\bar{Y}](R) \cap \pi[\bar{Y}](S)$
- c) (falls Anfrageoptimierung und interne Auswertung in der Vorlesung bereits besprochen wurden):
wie werden die jeweiligen Ausdrücke am effizientesten ausgewertet?