Datenbanksysteme WS 2016/17
Prof. Dr.-Ing. Sebastian Michel
BSc. Peter Brucker
TU Kaiserslautern, FB Informatik – Lehrgebiet Informationssysteme
http://dbis.informatik.uni-kl.de
Übungsblatt 5: Ausgabe 29.11.2016, Abgabe 05.12.2016

Aufgabe 1: Verschachtelte Anfragen (1 P.)

(a) Betrachten Sie folgenden Ausschnitt des TPC-H Schemas:

```
CREATE TABLE partsupp(ps_availqty INT, ps_partkey INT);
CREATE TABLE lineitem(l_partkey INT, l_shipdate DATE);
```

Darauf wird folgende Anfrage ausgeführt:

```
SELECT ps_partkey
FROM partsupp
WHERE ps_availqty = (SELECT COUNT(l_shipdate)
    FROM lineitem
    WHERE ps_partkey = l_partkey
    AND l_shipdate < '1-1-1999')
```

Betrachten Sie nun die folgende Entschachtelung.

```
WITH temp AS (
    SELECT l_partkey partkey, COUNT(l_shipdate) counted
    FROM lineitem
    WHERE l_shipdate < '1-1-1999'
    GROUP BY l_partkey
)
SELECT ps_partkey
FROM partsupp, temp t
WHERE ps_availqty = t.counted AND ps_partkey = t.partkey
```

- Zeigen Sie anhand von Beispieldaten, dass diese Entschachtelung nicht korrekt ist.
- Geben Sie anschließend eine korrekte Entschachtelung an.

(b) Betrachten Sie die folgenden beiden Möglichkeiten die Namen von Mitarbeitern zu berechnen, die mehr als 100 000 (Euro) verdienen und genauso alt wie der Manager ihrer Abteilung sind.

Zuerst, eine verschachtelte Anfrage

```
SELECT  e1.ename
FROM Employees e1
WHERE e1.salary > 100000
        AND e1.age = (SELECT e2.age
        FROM Employee e2, Department d2
        WHERE e1.dname = d2.dname
                AND d2.manager = e2.ename)
```

Zweitens, eine Anfrage, die eine View benutzt

```
SELECT e1.ename
FROM Employees e1, ManagerAge a
WHERE e1.dname = a.dname AND e1.salary > 100000 AND e1.age=a.age
```

```
CREATE VIEW ManagerAge(dname, age) AS
        SELECT d.dname, e.age
        FROM Employee e, Department d
        WHERE d.manager = e.ename
```

(i) Beschreiben Sie eine Situation, in der die erste Anfrage voraussichtlich günstiger als die zweite Anfrage ist.

(ii) Beschreiben Sie eine Situation, in der die zweite Anfrage voraussichtlich günstiger als die zweite Anfrage ist.

(iii) Können Sie eine äquivalente Anfrage erstellen, die effizienter als beide oben genannten Möglichkeiten ist, wenn jeder Angestellte, der mehr als 100 000 (Euro) verdient entweder 35 oder 40 Jahre alt ist? Beschreiben Sie Ihre Lösung/Antwort.

Aufgabe 2: Materialized Views (1 P.)

(a) **Wartung von Aggregationen**

Gegeben folgende Definition einer Materialized View über die Relation T mit $Sch(T) = [a, b]$:

```
CREATE MATERIALIZED VIEW V AS
SELECT a, AGG(b) FROM T GROUP BY a
```

Beschreiben Sie in der folgenden Tabelle den Aufwand, der bei einer Einfügung und Löschung eines Tupels (a,b) in T entsteht, für unterschiedliche Wahl der Aggregation `AGG`:

Aggregation	insert (a,b)	delete (a,b)
`COUNT`	Zähler für die Gruppe a muss inkrementiert oder neu angelegt und mit 1 initialisiert werden, falls a nocht nicht vorhanden war.	Zähler für a muss dekrementiert werden und gelöscht werden falls er bei 0 angekommen ist.
`SUM`		
`AVG`		
`MIN`		
`MAX`		

(b) **Wartung von Selektionen und Joins**

Gegeben Sichten $V_1 = \sigma_\theta(R)$ und $V_2 = R \bowtie S$. Geben Sie wie in der Vorlesung in relationaler Algebra an, wie sich V_1 und V_2 ändern, wenn Sie

- eine Menge an Tupeln i_r in R einfügen,
- eine Menge an Tupeln d_r aus R löschen.

(c) **Wartung von Projektionen**

Beschreiben Sie die Herausforderung bei der Wartung einer View $V_3 = \pi_a(R)$, wobei $Sch(R) = [a, b]$. Wie kann diese View trotzdem relativ effizient gewartet werden?

Aufgabe 3: Denormalisierung, Partitionierung, Rewriting (1 P.)

(a) Betrachten Sie das Schema aus Aufgabe 3 von Blatt 1 erneut für die folgende Teilaufgabe.

Der Autohändler Karl-Heinz Müller beschwert sich über die schlechte Performance seiner Datenbank-gestützten Verwaltungssoftware. Folgende Relationen sind vorhanden:

Kunden: [KuNr, Vorname, Name, Straße, PLZ, Ort]
verkauft: [KuNr, KfzNr, VerkäuferPersNr, Preis, Datum]
Autos: [KfzNr, Hersteller]
Mitarbeiter: [PersNr, Vorname, Nachname, Telefon]

Wie können Sie das bekannte Schema abändern, um die folgenden Anfragen sehr effizient zu beantworten? Welche Trade-offs handeln Sie sich dadurch ein?
Hinweis: Berücksichtigen und beschreiben Sie in Ihrer Antwort auch die Normalformen des originalen und abgeänderten Schemas.

- A5: Der Hersteller, mit dem am meisten Umsatz gemacht wurde.
- A6: Der Ort, aus dem die meisten Käufer kommen.

(b) Verwenden Sie ein Beispiel aus dem Uni-Schema um die Begriffe **vertikale** und **horizontale Partitionierung** zu beschreiben. Stellen Sie weiterhin jeweils Vor- und Nachteile bzw. Probleme der beiden Partitionierungsmöglichkeiten gegenüber.

(c) Diskutieren Sie für die unten stehenden Anfragen auf das bekannte Uni-Schema evtl. auftretende Schwierigkeiten für den Anfrageoptimierer. Geben Sie jeweils eine effizientere, ergebnisäquivalente Formulierung der Anfrage an. Beachten Sie die Angaben zu evtl. vorhandenen Indexen oder anderen Besonderheiten.

(i) Es gibt keinen Index.

```
SELECT max(Semester)
FROM Studenten S
GROUP BY S.MatrNr
HAVING S.MatrNr=4121
```

(ii) Es gibt einen Index auf *Fachgebiet* und einen auf *Boss*.

```
SELECT *
FROM Assistenten A
WHERE A.Fachgebiet = 'Datenbanksysteme' OR A.Boss=2127
```

(iii) Es gibt keinen Index.

```
SELECT DISTINCT *
FROM Studenten S
```

(iv) Attribut *gelesenVon* in Vorlesungen ist Fremdschlüssel, der auf *PersNr* in Professoren verweist.

```
SELECT P.PersNr
FROM Professoren P, Vorlesungen V
WHERE  P.PersNr = V.gelesenVon
```

(v) Es gibt einen B^+-Index auf *alter*.

```
SELECT S.MatrNr
FROM Studenten S
WHERE 2*S.alter<40
```

Aufgabe 4: Dies und Das (0 P.)

- Wieso ist es sinnvoll, eine höhere Anzahl von Merge-Phasen beim Blocked-I/O des externen Sortierens in Kauf zu nehmen?
- Ist der Sort-Merge-Join generell günstiger als der Blocked-Nested-Loop Join? Begründen Sie Ihre Antwort.
- Geben Sie einen Join zwischen zwei Relationen $R(A, B)$ und $S(B, C)$ an, der via Sort-Merge-Join berechnet werden kann und einen über diesen Relationen, der nicht via Sort-Merge-Join berechnet werden kann.
- Beschreiben Sie zwei Möglichkeiten Duplikateliminierung zu implementieren.
- Wie ist die Selektivität eines Joins definiert?
- Eine Seite hat 4KB. Die Festplatte hat eine Latenz von 10ms und eine Lesegeschwindigkeit von 100MB/s. Geben Sie eine Formel für die Lesekosten in Sekunden an für externes Sortieren mit Blocked-IO von Parameter b für eine Datei mit N Seiten und n_B verfügbaren Seiten im Hauptspeicher.
- Beschreiben Sie was bei FM-Sketch und KMV-Sketch schief gehen kann, wenn die Daten anfangs nicht mit einer Hashfunktion abgebildet werden. Gehen Sie dabei konkret auf Probleme in der Funktionsweise der beiden Sketches ein.
- Geben Sie die Definition von C_{out} an. Wieso werden die Kosten für die einzelnen Relationen nicht berücksichtigt?
- Gegeben ein Join-Baum T. Geben Sie die Berechnungsvorschrift für die Kardinalität des Anfrageergebnisses an.
- Geben Sie ein Beispiel mit drei Relationen an, für welches der Join, der ein Kreuzprodukt enthält, am günstigsten ist.
- Eine Datei habe die Größe von N Blöcken und es seien n_B viele Blöcke im Hauptspeicher verfügbar. Wie viele Runs werden initial erzeugt und wie viele Durchläufe benötigt ein 2-way Merge bzw. ein $(n_B - 1)$-way Merge?
- Beschreiben Sie, wie durch Hashing Duplikate entfernt werden können.
- Was können Sie über die Kosten von $((S \bowtie R) \bowtie T) \bowtie U$ gegenüber $((S \bowtie T) \bowtie R) \bowtie U$ sagen, wenn Sie wissen, dass $(S \bowtie T) \bowtie R$ der optimale Baum für $\{R, S, T\}$ ist?
- Wie viele zigzag Bäume gibt es für n Relationen, wenn Kreuzprodukte zugelassen werden?
- Welche ursprüngliche Werte werden durch die Wavelet-Transformierte $[4, 2, 1, 0]$ beschrieben?
- Nennen Sie zwei Anwendungen von Sortierung in Datenbanksystemen.
- Wie lautet die Berechnungsvorschrift des optimalen Fehlers für i Datenpunkte und k $BucketsSSE^*(i, k)$ für V-Optimale Histogramme?
- Wieso kann es in Anfragegraphen, die die Form eines Sterns haben, keine buschigen Bäume geben?
- Gegeben eine Zufallsvariable X mit Verteilungsfunktion F. Was ist das p Quantil von X?
- Die Verteilung der Attributausprägungen für Attribut A seien beschrieben durch die Normalverteilung $\mathcal{N}(100, 5)$. Die Tabelle R habe n Tupel. Wie viele Tupel qualifizieren sich voraussichtlich für die Anfrage $\sigma_{A>100}(R)$.
- Beschreiben Sie das Optimalitätsprinzip für das Join-Ordering Problem.
- Welche Kriterien müssen erfüllt sein, damit eine Anfrage eine vorhanden View verwenden kann?
- Wann könnte es Sinn machen ein Relationenschema in 3NF Relation einem Relationenschema in BCNF vorzuziehen?

- Wie viele links-tiefe Bäume gibt es für n Relationen, wenn Kreuzprodukte zugelassen werden?
- Beschreiben Sie, wie durch Sortierung Duplikate entfernt werden können.
- Wie sind zigzag Bäume definiert?
- Was ist ein dependent Join?
- Was gilt es zu beachten, wenn mehrere Anfragen gleichzeitig optimiert werden sollen?
- Wie viele links-tiefe Bäume sind bei Stern-Anfragen mit n Relationen möglich, wenn keine Kreuzprodukte zugelassen werden?
- Welchen Schätzwert können Sie für einen Hash-Sketch der Form 110101011011 ablesen?
- Wann nennen wir eine Hashfunktion h (r_1, r_2, p_1, p_2)-sentitive?
- Kann ein V-Optimales Histogramm mit 4 Zellen besser sein als ein Equi-Width-Histogramm mit 8 Zellen?
- Beschreiben Sie den Einfluss der Anzahl von Hashfunktionen auf Precision und Recall in LSH.