

Aufgabenblatt 4 Anfrageausführung

- Abgabetermin: **Mittwoch, 12.1.2011 (23:59 Uhr)**
- Zur Prüfungszulassung muss ein Aufgabenblatt mit mind. 25% der Punkte bewertet werden und alle weiteren Aufgabenblätter mit mindestens 50% der Punkte.
- Die Aufgaben sollen in Zweiergruppen bearbeitet werden.
- Abgabe:
 - per E-Mail an `dbs2-2010@hpi.uni-potsdam.de` mit Subject
Abgabe DBS II: Aufgabenblatt <n> Namen
 - ausschließlich pdf-Dateien
 - eine Datei pro Aufgabe mit folgendem Dateinamen:
`blatt<aufgabenblattNr>aufgabe<aufgabenNr><Nachnamen>.pdf`
Bitte **keine Leerzeichen, Unterstriche, Umlaute, Sonderzeichen, ...** im Dateinamen!
 - **jedes Blatt beschriftet mit Namen**
 - Wir korrigieren die Abgaben aufgabenweise. Das beschriebene Verfahren vereinfacht uns die Arbeit erheblich!

Aufgabe 1: Block-basierter Nested-Loop-Join

Gegeben seien zwei Relationen S und R . Die Tupel der Relationen umfassen jeweils 10.000 Blöcke: $B(S) = B(R) = 10.000$. Der verfügbare Hauptspeicher umfasst $M = 1.001$ Blöcke.

- a) Berechne die Anzahl an I/O-Operationen für die Berechnung von $R \bowtie S$, wenn ein Block-basierter Nested-Loop-Join-Algorithmus verwendet wird. **3 P**
- b) Wie groß muss M mindestens sein, wenn $R \bowtie S$ unter Verwendung des Block-basierten Nested-Loop-Join-Algorithmus mit nicht mehr als 100.000 I/O-Operationen berechnet werden soll? **4 P**
- c) Wie groß muss M mindestens sein, wenn $R \bowtie S$ unter Verwendung des Block-basierten Nested-Loop-Join-Algorithmus mit nicht mehr als 25.000 I/O-Operationen berechnet werden soll? **3 P**

Aufgabe 2: Sort-Merge Join-Algorithmus

Gegeben seien zwei Relationen $R(\underline{A}, B)$ und $S(\underline{A}, C)$. Die Tupel in R umfassen insgesamt $2^{13} = 8.192$ Blöcke und die Tupel in S belegen insgesamt $2^{10} = 1.024$ Blöcke. Für die Berechnung von $R \bowtie S$ können insgesamt $2^7 = 128$ Blöcke des Hauptspeichers genutzt werden.

- a) Wie viele sortierte Teillisten entstehen in der ersten Phase des Two-Phase Multiway Merge-Sort (TPMMS) bei der Sortierung von R ? Wie viele für S ? **2 P**
- b) Wie viele Blöcke müsste eine der beiden Eingaberelationen *mindestens* umfassen, damit die Sortierung mittels des TPMMS *nicht* mehr möglich ist? **2 P**
- c) Notiere kurz die Ausführungsschritte des Sort-Merge Join-Algorithmus für die Berechnung von $R \bowtie S$. Notiere zusätzlich für jeden Teilschritt die Anzahl an I/O-Operationen.
Hinweis : Beachte, dass das Attribut A in beiden Relationen Primärschlüssel ist. **4 P**
- d) Angenommen die Blockanzahl von S vergrößert sich auf $2^{13} + 1 = 8.193$ Blöcke. Welches Problem ergibt sich dann bei der Berechnung von $R \bowtie S$ unter Verwendung des Sort-Merge Join-Algorithmus? Begründe deine Antwort. **2 P**

Aufgabe 3: Sort-basierte Two-Pass Algorithmen

Gegeben seien zwei Relationen $R(X, Y)$ und $S(Y, Z)$. Die Tupel in R umfassen insgesamt 1000 Blöcke und die Tupel in S belegen insgesamt 500 Blöcke. Berechne zwei Dinge, um die folgenden Sort-basierten Two-Pass Algorithmen durchführen zu können: Die nötige Blockanzahl im Hauptspeicher und die I/O-Kosten - *ohne* das Schreiben des Ergebnisses auf die Platte:

- a) Mengen-Vereinigung 2 P
- b) Einfacher, Sort-basierter Join. 2 P
- c) Verbesserter, Sort-basierter Join. Notiere zusätzlich für diese Join-Variante kurz die notwendigen Annahmen. 3 P

Ermittle nun die Anzahl an I/O-Operationen für die Bestimmung von $R \bowtie S$ unter Verwendung des einfachen Sort-basierten Join-Algorithmus, wenn es nur zwei unterschiedliche Y -Werte gibt: $V(R, Y) = V(S, Y) = 2$. Jeder Y -Wert tritt dabei in der Hälfte der Tupel aus R und der Hälfte der Tupel aus S auf. Für die Berechnung können insgesamt 101 Blöcke des Hauptspeichers genutzt werden. 4 P

Aufgabe 4: Hashjoin

Gegeben seien zwei Relationen R und S . Die Tupel in R umfassen insgesamt 400 Blöcke und die Tupel in S belegen insgesamt 500 Blöcke. Für die Berechnung von $R \bowtie S$ können insgesamt 40 Blöcke des Hauptspeichers genutzt werden.

- a) Um I/O-Kosten zu sparen, wird während des Hashjoins 1 Bucket komplett im Speicher behalten. Ermittle die Anzahl der Buckets, die benötigt werden, um den Hashjoin weiterhin in zwei Phasen berechnen zu können, unter der Annahme, dass alle Buckets dieselbe Größe haben. 6 P
- b) Berechne die Anzahl an I/O-Operationen für das in Teilaufgabe a) beschriebene Vorgehen zur Bestimmung von $R \bowtie S$. Verwende in deiner Rechnung k als die in Teilaufgabe a) ermittelte Anzahl der benötigten Buckets. 3 P