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 1: Ausgabe 25.10.2015, Abgabe 07.11.2015

Aufgabe 1: Seitenersetzungsstrategien (1 P.)

Gegeben folgende Sequenz von Seitenzugriffen:

C, B, D, D, C, E, B, C, E, A, A, E, A, A, A

(a) Berechnen Sie für diese Sequenz bei einer Puffergröße von 3 die Anzahl von Zugriffen auf Seiten, die sich zu dem Zeitpunkt nicht im Puffer befinden, für die Strategien CLOCK, LRU-k mit $k = 2$ und GCLOCK.
Gehen Sie für GCLOCK von einem Initialwert $E_i = 3$ für jede Seite und einem Inkrementwert bei Referenz $W_i = 1$ aus.

(b) Welche Ersetzungen finden für diese Sequenz bei einer optimalen Strategie, die zu jedem Zeitpunkt bereits alle zukünftigen Seitenzugriffe kennt[1], statt?

(c) Finden Sie ein Beispiel für eine Seitenzugriffssequenz, bei dem die Strategie FIFO echt weniger Cachemisses hat als LFU, und ein Beispiel bei dem FIFO echt weniger Hits hat als LFU. Nehmen Sie eine Puffergröße von 2 an.

Aufgabe 2: Performance-Überlegungen mit PostgreSQL und pgAdmin (1 P.)

In diesem und den folgenden Übungsblättern werden wir das relationale Datenbanksystem *PostgreSQL* benutzen. PostgreSQL ist unter einer Open-Source-Lizenz und somit kostenlos verfügbar, etwa unter `http://www.postgresql.org/` als installierbare Pakete oder für Linux-Distributionen über die gängigen Paketmanager (`apt-get, yum, pacman, ...`). Installieren Sie sich PostgreSQL (Version 9.3 oder höher) auf einem Ihrem System angemessenem Weg. Zusätzlich gibt es mit *pgAdmin* eine graphische Benutzeroberfläche, die die Benutzung und Administration von PostgreSQL vereinfacht. Installieren Sie auch dieses Programm (`http://www.pgadmin.org/` oder Paketmanager). Sollten Sie wider Erwarten Probleme mit der Installation haben, melden Sie sich frühzeitig(!) per Email unter Angabe Ihres SCI-Benutzernamens an mhoffmann@cs.uni-kl.de, damit Sie einen Zugang zu einer PostgreSQL-Installation im SCI bekommen.

Laden Sie die beiden in der Vorlesung vorgestellten Datensätze in die Datenbank. Nutzen Sie die Gelegenheit, um Ihre SQL-Kenntnisse aufzufrischen und ein paar Anfragen gegen die Datenbank stellen.

Hinweis: Wenn pgAdmin Ihnen Popups anzeigt, in denen z.B. ein VACUUM-Lauf empfohlen wird, ignorieren Sie das. Diese Tips sind vor allem für Datenbankadministratoren von Produktivsystemen gedacht.

Effekte von Indizes

Sie können den Anfrageplan, der vom Datenbanksystem generiert wird, ausgeben lassen, indem Sie "Explain" (oder "Analysieren") wählen.

(a) Welcher Anfrageplan wird für folgende Anfrage generiert?

```
SELECT * FROM lineitem WHERE l_orderkey < 10;
```

[1] Solch eine Strategie (clairvoyant) kann natürlich nicht implementiert werden. Allerdings wird diese Art der Betrachtung häufig benutzt, um die Qualität anderer Verfahren zu bewerten.

(b) Was ändert sich am Anfrageplan, wenn Sie zuerst den folgenden Index erzeugen lassen?

```
CREATE INDEX index1 ON lineitem(l_orderkey);
```

(c) Die SELECT-Anfrage aus Aufgabenteil (a) liefert alle Einträge mit `l_orderkey` < $k = 10$. Ab welcher Größe von k ändert sich der Anfrageplan? Wie verhält sich dieses k zur Gesamtzahl der Einträge der `lineitem`-Tabelle?

(d) Wie sieht der Anfrageplan für die folgende Query aus und was ist der Unterschied zu zuvor?

```
SELECT l_orderkey FROM lineitem WHERE l_orderkey < 10;
```

Aufgabe 3: Indextuning (1 P.)

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]

Laut Herrn Müller müssen insbesondere folgende Anfragen häufig ausgeführt werden:

- A1: Autos eines bestimmten Herstellers müssen unbedingt in die Werkstatt. Nicht alle, sondern nur die, mit Kaufdatum später als X.
- A2: Die Gesamtumsätze (Summe der Verkaufspreise) zu jedem Hersteller.
- A3: Die Liste der 100 zuletzt verkauften Autos zusammen mit dem Käufer-Namen, geordnet nach Kaufdatum.
- A4: Die Telefonnummern aller Mitarbeiter.

Welche Indexe sollten angelegt werden, um die Performance zu verbessern? Diskutieren Sie Sinn und Zweck der Indexe, evtl. können Indexe gleich in mehreren Anfragen genutzt werden. Die Primärindexe sind bereits vorhanden. Bei welchem dieser Indexe bietet sich als Alternative zum B+-Baum ein Hash-Index an?

Aufgabe 4: InSy in a Nutshell (0 P.)

Diese Aufgabe befasst sich mit Inhalten, die Ihnen aus der Vorlesung Informationssysteme (Voraussetzung für Datenbanksysteme) bekannt sein sollten.

(a) Nennen und erklären Sie die Bedeutung der Buchstaben A, C, I und D in der Abkürzung ACID.

(b) Erklären Sie die beiden Phasen des 2PL-Protokolls.

(c) Was bedeutet Verlustlosigkeit bei der Zerlegung einer Relation R in Relationen R_1 und R_2 und welche hinreichende Bedingung gibt es dafür?

(d) Gegeben eine Relation `Universitaet(Stadt, Studenten, Gruendungsjahr)`. Geben Sie eine Fensteranfrage unter Verwendung von `rank()` und einer entsprechenden Fensterdefinition an, die Folgendes findet: Die Namen aller Städte, deren Universität in ihrem Gründungsjahr zu den drei populärsten Universitäten zählt, was die Anzahl an Studenten betrifft.

(e) Drücken Sie $R \ltimes_\theta S$ allein durch Kreuzprodukt, Selektion, Projektion und Mengenoperationen aus. Die Attribute einer Relation R können Sie in Ihrer Lösung als $Attr(R)$ bezeichnen.
Erklären Sie weiterhin kurz die Begriffe „Äußerer Join“ und „Anti-Join“ sowie den Unterschied zwischen „Natürlichem Join“ und „Equi-Join“.

(f) Bringen Sie die Relation R mit Schema $\mathcal{R} = \{A, B, C, D, E, F\}$ und FDs $\mathcal{F}_\mathcal{R} = \{A \to BC, CE \to BF, BF \to A\}$ in 3NF.

(g) Gegeben zwei Relationen R_1 und R_2, mit $|R_1| = n$ und $|R_2| = m$, geben Sie für folgende Ausdrücke der relationalen Algebra jeweils die minimale und die maximale Anzahl an Ergebnissen an. Gehen Sie von Mengensemantik aus.

- $R_1 \ltimes R_2$
- $R_1 \times R_2$
- $\sigma_P(R_1)$, wobei P ein beliebiges Prädikat ist
- $\pi_A(R_2)$, wobei A eine beliebige Menge von Attributen aus dem Schema von R_2 ist

(h) Histogramme: Folgende Integer-Werte aus $[0, 19]$ seien gegeben:

2, 7, 9, 8, 4, 5, 4, 9, 8, 10, 15, 15, 19, 18, 1, 0, 1, 2

Fügen Sie diese Werte in ein Equi-Width-Histogramm mit Zellenbreite 4 ein. Geben Sie für jede Zelle deren Grenzen sowie Häufigkeit an.
Hinweis: Sie müssen das Histogramm nicht unbedingt zeichnen.

Berechnen Sie anhand dieses Histogramms für folgende Anfragen die geschätzte Kardinalität sowie den absoluten Fehler der Schätzung gegenüber des tatsächlichen Wertes.

- Wie oft tritt der Wert 8 auf?
- Wie viele Werte liegen im Bereich $[5, 12]$?
- Wie viele Werte liegen im Bereich $[1, 8] \cup [12, 17]$?