Aufgabenblatt 3
Indexstrukturen und Anfrageausführung

- Abgabetermin: **Dienstag, 8.12.09 (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 dbs2-200910@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!

- **Die DB2-Dokumentation findest du unter:**
  http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp
- Bitte bearbeite die Aufgaben in der vorgegebenen Reihenfolge, so dass die Ergebnisse nachvollziehbar bleiben.
- Gib zu jeder Aufgabe die geforderten Anfragebäume mit an.

Aufgabe 1: Erzeugen der Datenbank und Laden der Daten

- Erzeuge in deiner Instanz eine Datenbank unter Verwendung der Anweisungen in create_database.sql (im Homeverzeichnis deines Instanznutzers). Die Tablespaces werden im Homeverzeichnis deines Instanznutzers angelegt. Ersetze die Angaben durch deine Instanzwerte.
- Erzeuge das TPC-H Schema mit dem Skript create_tpch_schema.sql.
- Lade die Daten, setze die Integrität und sammle die Statistiken unter Verwendung der Anweisungen in load_tpch_data.sh. Hinweis: Miss die Zeit zum Laden der Daten in die customer-Tabelle (benötigt in Aufgabe 3).

Generelle Hinweise:

- Server dbs2-200910; Zugriff per HPI-Account
- Jede Gruppe hat eigene Instanz. Die Zuordnung der Instanzen findet ihr unter
  lehrveranstaltungen\DBSII_naumann\zuordnungInstanzen.pdf
- Anmeldung als Instanz-Eigner per dbs2login <instanz>
- Die Skripte liegen im Homeverzeichnis der Instanz.
  (Ausführen von sql-Skripten per db2 -tf <skript.sql>)
- Die TPC-H-Daten liegen unter /mnt/tpchData-1G/tables/
- Starten der Instanz per db2start; Stoppen per db2stop

Aufgabe 2: Indexe und Workloads

Betrachte die Relation Customer und ihre Indexe. Erstelle einen weiteren, sortierten Index auf dem Attribut c_acctbal und sammle alle Statistikinformationen.

Der Workload sei wie folgt gegeben:

Q1: **select** $*$ **from** TPCH.customer **where** c_custkey < 50000

Q2: **select** $*$ **from** TPCH.customer **where** c_custkey < 100000

Q3: **select** $*$ **from** TPCH.customer **where** c_custkey < 140000

Q4: **select** $*$ **from** TPCH.customer **order by** c_acctbal

Q5: **select** $*$ **from** TPCH.customer **where** c_acctbal **between** 0 **and** 1000 **order by** c_acctbal

a) Betrachte die Ausführpläne zu Q1, Q2 und Q3.
   1) Gibt es einen echten Unterschied der Ausführpläne zu Q1 und Q2? Warum (nicht)? **4 P**
   2) Was ist der Unterschied zwischen den Ausführplänen zu Q2 und Q3? Warum werden deiner Meinung nach die Pläne so gewählt? **4 P**

b) Betrachte die Ausführpläne zu Q4 und Q5. Was ist der Unterschied? Warum werden deiner Meinung nach die Pläne so gewählt? **4 P**

Aufgabe 3: Import vs. Load

Es gibt zwei Varianten Daten in eine DB2-Datenbank zu bringen – Import und Load. Per Load haben wir bereits alle Daten in das Schema geladen.

Lege nun eine zweite, leere customer-Tabelle an und *importiere* die Daten in diese Tabelle.

a) Vergleiche die Laufzeiten für beide Varianten. **4 P**

b) Wie erklärst du dir diesen Unterschied? Nutze die DB2-Doku für deine Antwort. **4 P**

Aufgabe 4: Mehrattributige Indexe I (Verwendung von Zusatzattributen)

Betrachte folgende Anfragen und deren Ausführpläne mit folgenden Indexen:

Anfragen

Q1: **select** o_orderkey **from** TPCH.orders **where** o_orderkey > 100000

Q2: **select** o_orderkey, o_custkey **from** TPCH.orders **where** o_orderkey > 100000;

Q3: **select** o_orderkey, o_custkey, o_orderstatus **from** TPCH.orders **where** o_orderkey > 100000;

Q4: **select** o_orderkey, o_custkey, o_orderstatus, o_totalprice **from** TPCH.orders **where** o_orderkey > 100000;

Indexe

I1: **create unique index** orders_idx1 **on** orders (o_orderkey) include (o_custkey) collect detailed statistics

I2: **create unique index** orders_idx2 **on** orders (o_orderkey) include (o_custkey, o_orderstatus) collect detailed statistics

a) Welche Ausführpläne werden für die Anfragen gewählt

   1) mit Index I1
   2) mit Index I1 und I2.

   Erkläre die Wahl der Indexe. **6 P**

b) Wird ein Index verwendet für Q4? Wie müsste man die Anfrage verändern, damit ein Index verwendet wird? (2 Möglichkeiten) Begründe deine Antwort. **4 P**

Aufgabe 5: Mehrattributige Indexe II

a) Lege den folgenden Index an:

   create index orders_idx3 **on** orders (o_orderkey, o_custkey) collect detailed statistics

   Was passiert? Was sagt dir das über Indexe mit Zusatzattributen in DB2? **4 P**

b) Betrachte die Ausführpläne für die folgenden Anfragen.

   Q5: **select** o_orderkey, o_custkey **from** TPCH.orders **where** o_orderkey > 100000

   Q6: **select** o_orderkey, o_custkey **from** TPCH.orders **where** o_custkey > 100000

   Werden Indexe verwendet? Was sagt dir das über multidimensionale Indexe in DB2? **4 P**

c) Lege den folgenden Index an:

   create index orders_idx4 **on** orders (o_custkey, o_orderkey) collect detailed statistics

   Verändert sich der Ausführplan für Q6? Was sind die Unterschiede? Was sagt dir das über multidimensionale Indexe in DB2? **4 P**