Aufgabenblatt 4
Benchmarking

- Abgabetermin: **Dienstag, 12.01.2015 (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 Aufgabe soll in Zweiergruppen bearbeitet werden.
- Abgabesytem unter
  `http://www.dcl.hpi.uni-potsdam.de/submit`
  - ausschließlich pdf-Dateien
  - eine Datei namens Aufgabe-1.pdf
  - beschriftet mit Namen
- **Die DB2-Dokumentation findest du unter:**
  `http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/`

Aufgabe 1: TPC-H Power Test

Gegeben ist eine virtuelle Maschine (VM), in der eine DB2-Express Datenbank läuft. In die Datenbank wurde bereits ein 1GB Datensatz des TPC-H eingelesen. Wir wollen diesen Datensatz nutzen um eine Reihe von TPC-H Anfragen zu beantworten. Ziel dieses Aufgabenblattes ist es, die gegebene Datenbank so zu optimieren, dass ein gegebener Workload möglichst schnell abgearbeitet werden kann.

Hintergrund zum TPC-H Benchmark
Im Rahmen des TPC-H Benchmarks werden zunächst neue Tupel in die Tabellen *lineitem* und *orders* eingefügt, anschließend 22 Anfragen abgesetzt und zuletzt Daten aus den genannten Tabellen gelöscht. **In dieser Übung verwenden wir nur die 22 Anfragen ohne Inserts und Deletes!** (Wer will, kann diese zum Selbststudium aber natürlich auch noch betrachten)

Anforderungen zum Bestehen des Aufgabenblattes
Optimiere die TPC-H Datenbank, um die Ausführungszeiten der TPC-H Anfragen zu beschleunigen (z.B. über geeignete Indexe). Dokumentiere und begründe alle ergriffenen Maßnahmen:

- Was für ein Rechner wurde für die Tests genutzt (falls kein Poolrechner genutzt wurde)?
- Warum wurde ein bestimmter Index angelegt und warum andere Indexe evtl. verworfen?
- Wurden nicht-indexbasierte Optimierungen in Erwägung gezogen und wenn ja, welche?
- Wie haben sich die Ausführungszeiten der einzelnen Queries und die Gesamtausführungszeit durch die Anpassungen geändert? Sammelt eure Ergebnisse für diese Diskussion tabellarisch in der Form *(Query [#], Time_original [sec], Time_optimized [sec], Reduction [%])*. Ein Eintrag soll in dieser Tabelle auch für die Gesamtlaufzeit aufgeführt werden.

Das Aufgabenblatt gilt als bestanden, wenn die (ergriffenen oder verworfenen) Maßnahmen nachvollziehbar dokumentiert und begründet sind. Es sollten mindestens 10 Maßnahmen in Erwägung gezogen werden! Fasse deine Änderungen an den TPC-H Tabellen außerdem in einem ausführbaren SQL-Skript (Aufgabe-1.sql) zusammen und **füge dessen Inhalt der Dokumentation (Aufgabe-1.pdf) an**.

Spielregeln
Das Skript *db-config.sh* (siehe unten) **muss** vor der Bearbeitung des Aufgabenblattes ausgeführt werden. Es setzt die wichtigen DB2-Variablen auf vorgegebene Werte. Diese Werte dürfen zur Optimierung *nicht* geändert werden! Es handelt sich dabei unter anderem um folgende Werte:

- Die maximale Größe des userspace1, die auf 4 GB gesetzt wird.
- Die Datenbankparameter AUTO_MAINT und SELF_TUNING_MEM, die auf OFF gesetzt werden um die Auswirkungen individueller Optimierungsschritte nachvollziehen zu können.

Es dürfen außerdem keine Materialisierten Sichten erstellt werden!

Allgemeine Hinweise zur VM
Für die praktischen Aufgaben stellen wir die virtuelle Maschine "DB2 Express-C 9.7 32-bit" zur Verfügung. Die virtuelle Maschine (VM) ist folgendermaßen eingerichtet:

- Betriebssystem: *SUSE Linux Enterprise Server*
- Datenbank: *DB2 Express-C 9.7*
- Datenbank-Instanz: *db2inst1*
- Datenbank-Name: *db2db1* (enthält bereits alle Tabellen und Daten des TPC-H mit scale factor 1, also 1GB)

In der virtuellen Maschine wurde ein Nutzer für das Betriebssystem und die Datenbank angelegt mit den folgenden Zugangsdaten:

- Nutzername: *db2inst1*
- Passwort: *ws2011*

Hinweise zum Umgang mit der VM

- Arbeit an einem Poolrechner:
  - Start der VM über:
    `Windows > All Programs > VMware > DB2 Express-C 9.7 32-bit`
- Arbeit mit dem eigenen Rechner:
  - Voraussetzung: VMware Player (Freeware)
  - VM kopieren von Laufwerk
    `R:\lehrveranstaltungen\FG_Informationssysteme\VL DBS II\uebung benchmarking`
    - Alle Dateien werden benötigt
    - Die ausführbare Datei ist *DB2 Express-C 9.7 32-bit.vmx*
  - Starten der VM über:
    - Starte VMware Player > Open a Virtual Machine > *DB2 Express-C 9.7 32-bit.vmx*
    - ODER: Doppelklick auf *DB2 Express-C 9.7 32-bit.vmx*
- Die virtuellen Festplatten der VM sind nicht schreibbar, d.h. alle Änderungen an einer laufenden Instanz gehen nach dem Neutstart der VM verloren, so dass sich die VM wieder im „Auslieferungszustand" befindet. Falls du die VM auf deinem eigenen Rechner verwendest, besteht jedoch die Möglichkeit sie im VMware Player zu „suspenden" (Virtual Machine > Power > Suspend).
- **Auf den Poolrechnern darf die VM nicht suspended werden**, da du sie sonst für deine Kommilitonen auf dem Rechner blockierst. Daher muss die VM immer heruntergefahren werden (Computer > Shutdown).
- Beim Start fragt die VM evtl. ob Software-Updates installiert werden sollen. Das ist nicht notwendig (Remind me later).
- Einstellung für deutsches Tastaturlayout:
  `Computer > Control Center > Hardware > Keyboard > Layouts > Add > Germany > set as Default`
- Der Zugriff auf das Internet ist nicht freigegeben. Daten können aber per Drag-and-Drop in das VM-Fenster mit dem Host-Rechner ausgetauscht werden.

Möglichkeiten zum Ausführen von Anfragen

- Kommandozeile
  - Shell öffnen:
    `Computer > Gnome Terminal`
  - Verbindung mit der Datenbank herstellen:
    `db2 connect to DB2DB1`
  - Queries ausführen (Beachte die Anführungsstriche!):
    `db2 "<sqlquery>"`
  - Ein (selbstgeschriebenes) SQL-Skript ausführen:
    `db2 -tvf <skriptname>`
  - Verbindung mit der Datenbank trennen:
    `db2 connect reset`

- IBM Data Studio
  - *Data Studio* starten:
    `Desktop > DB2 Data Studio`
  - Verbindung zur Datenbank herstellen:
    `Administration Explorer > localhost/5001/DB2DB1 expandieren, rechtsklicken und "Connect" wählen > Nutzerdaten angeben > "Save Password" setzen > OK`
  - Projekt anlegen:
    `File > New > Data Development Project > Next > Finish > No`
  - Script anlegen:
    `Rechtsklick auf Projekt > New > SQL or SQuery Script > Finish`
  - SQL-Anfragen ausführen:
    - SQL in Script eintippen
    - SQL zum Ausführen markieren (falls nichts markiert ist, werden alle Anfragen ausgeführt)
    - Play-Button klicken (alternativ: Script > Run SQL)
  - Ausführungsplan anzeigen mittels Visual Explain:
    `SQL-Anfrage markieren > Rechtsklick > Open Visual Explain > Finish`

Zusätzliche Informationen und notwendige Skripte
Folgende Skripte werden benötigt:

- Das Shell-Skript *db-config.sh* setzt die notwendigen Parameter in der Datenbankinstanz. Führt dieses Skript einmal vor der Bearbeitung des Aufgabenblatts aus.
- Das SQL-Skript *runstats-tables.sql* aktualisiert die Statistiken im Systemkatalog der Datenbankinstanz. Diese Statistiken nutzt der Anfrageoptimierer zum Erstellen des Anfrageplans. Das Skript sollte also vor *jedem* Benchmark-Lauf einmal ausgeführt werden, damit die letzten Änderungen und Optimierungen im Benchmark angewandt werden können.
- Das Shell-Skript *runBenchmark.sh* führt den TPC-H Benchmark (ohne inserts und deletes) aus und misst gleichzeitig die Ausführungszeiten der einzelnen Anfragen. Die Ausführungszeit dieses Skriptes soll optimiert werden.
- Die 22 SQL-Skripte im Ordner *queries* sind die Anfragen, die im Benchmark eingesetzt werden.

Alle notwendigen Skripte, die VM und die TPC-H Spezifikation liegen auf dem HPI-Netzlaufwerk

`R:\lehrveranstaltungen\FG_Informationssysteme\VL DBS II\uebung benchmarking`

Hilfreiche Hinweise
Die TPC-H Daten sind bereits in die Datenbank db2db1 geladen und können daher sofort genutzt werden. Falls das Umsetzen von Optimierungen die Daten zerstören sollte, können die Daten folgendermaßen wieder hergestellt werden:

- Neuladen der Daten mit den Skripten aus dem Home-Verzeichnis des Nutzers oder
- einfach VM aus- und dann wieder anschalten um Initialzustand wiederherzustellen.

Im Home-Verzeichnis des Nutzers liegen außerdem

- optionale Updates und Skripte zum Zurücksetzen der Updates
- die Rohdaten und load logfiles
- die TPC-H queries (Achtung: query22 liegt hier in einer veralteten Version vor!)

Folgende Tools können bei der Bearbeitung der Aufgabe helfen:

- *Visual Explain* im DB2 Studio
- *db2advis* im Terminal

Falls db2advis den Fehler "Explain tables not set up properly ..." ausgibt, dann können die Explain Tables auf der Konsole folgendermaßen neu erstellt werden:

```
db2 -tf ~/sqllib/misc/EXPLAIN.DDL
```

In der gegebenen Datenbankinstanz existieren bereits Indizes. Jedoch könnten einige dieser Indizes für die Anfragebearbeitung sogar negative Folgen haben. Der Benchmark enthält nämlich keine Inserts und Deletes. Daher können bestehende Indexe die Anfragebeantwortung auch verlangsamen! Es sind daher möglicherweise nicht alle Hinweise von dbadvis sinnvoll Prüfe also, ob neue Indizes gesetzt oder auch vorhandene Indizes gelöscht werden sollten.

Die Verwaltung der Indexe ist die Hauptaufgabe dieser Übung. Es gibt noch kleine Kniffe zur Optimierung der Daten/Metadaten, die wir auch als Optimierung zählen würden. Ihr müsst (und sollt) aber KEINE Normalisierung oder Denormalisierung der Daten durchführen und KEINE materialisierten Sichten erstellen!

Um eindeutigere Messergebnisse zu erzielen bietet sich das mehrfache Ausführen des Benchmarks an. Ihr könnt dann die durchschnittlich, maximale und minimale Ausführungszeit zum Vergleich heranziehen. Der Caching-Einfluss sollte so minimiert werden.

Die korrekte, fachliche Diskussion einzelner Indexe ist uns wichtiger als der am Ende erzielte Performance-Gewinn.