

Übung DWH-DM

Auswertung Wettbewerb

Ulf Leser

Wissensmanagement in der
Bioinformatik



Erinnerung

- Drei Aufgaben zählten für den Wettbewerb
 - Berechnung von Differential Snapshots
 - Optimale Indexierung bei begrenzter Indexzahl
 - Optimale Materialized Views bei begrenzter Zahl und Bestrafung großer MVs

Ergebnisse

| | 1: Zeit | 1: Rang | 2: Zeit | 2: Rang | 3: Größe | 3: Zeit | 3: Score | 3: Rang |
|---------|---------|---------|---------|---------|-----------|---------|----------|---------|
| Gruppe1 | 10:45 | 2 | 0:65 | 5 | 1.543.952 | 285 | 0,43 | 3 |
| Gruppe2 | 20:47 | 4 | 11:00 | 8 | 1.916.962 | 294 | 0,49 | 5 |
| Gruppe3 | 27:82 | 7 | 1:65 | 6 | 36.503 | 774 | 0,60 | 6 |
| Gruppe4 | 22:27 | 5 | 0:40 | 3 | 1.142.135 | 385 | 0,45 | 4 |
| Gruppe5 | 50:13 | 8 | 0:44 | 4 | 2.108.240 | 858 | 0,96 | 8 |
| Gruppe6 | 26:16 | 6 | 0:39 | 2 | 500.396 | 215 | 0,23 | 1 |
| Gruppe7 | 17:90 | 3 | 8:50 | 7 | 33.755 | 444 | 0,35 | 2 |
| Gruppe8 | 10:01 | 1 | 0:15 | 1 | 2.392.511 | 348 | 0,60 | 6 |

Ergebnisse

| | 1: Punkte | 2: Punkte | 3: Punkte | Gesamt |
|---------|-----------|-----------|-----------|--------|
| Gruppe1 | 3 | 3 | 2 | 8 |
| Gruppe2 | | | | |
| Gruppe3 | | | | |
| Gruppe4 | | | | |
| Gruppe5 | | 5 | | 5 |
| Gruppe6 | | | 5 | 5 |
| Gruppe7 | 2 | | 3 | 5 |
| Gruppe8 | 5 | 2 | | 7 |

Mitnehmen: Benchmarking ist schwierig

- Cold cache / Hot cache
 - Riesige Unterschiede zwischen erster und zweiter Ausführung
 - Da alle auf eigenen Daten gearbeitet haben, hoffentlich wenig Cross-Gruppen Cache Effekte
- Externe Belastung des Rechners
- Ausnutzung mehrerer Prozessoren
 - Insbesondere bei Aufgabe 1
 - Auch bei Aufgabe 2/3 hätte man über Partitionierungen eine parallele Query-Verarbeitung versuchen können
- „Realistische“ Constraints
 - Warum nur 4 Indexe? 4 MV's?
 - Warum Größe bestrafen? Warum Laufzeit/Größe im Verhältnis 2:1?
- Was misst man?
 - Kommunikation zwischen Client und Server?
- Wie sähen die Ergebnisse auf andere Hardware aus?
- ...

Mitnehmen: Optimieren ist nur fast Wissenschaft

- Was macht dieser komische Optimierer?
- Sind Query Hints noch reine Lehre?
- Wie würde sich ein anderes System verhalten?
- Welche Statistiken soll man nehmen?
 - Wir haben das nur angekratzt (Histogramme, Sampling, Sampling Views, Self-Tuning, ...)
- Bestimmte Grundprinzipien sind aber überall gleich
- Database Tuning wäre eine eigene Vorlesung wert

Mitnehmen: Optimieren lohnt sich

- Verhältnis beste zu schlechtesten Zeiten
 - Aufgabe 1: 10:01 / 50:13 ~ Faktor 5
 - Aufgabe 2: 0:15 / 11:00 ~ Faktor 44
 - Aufgabe 3: 215 / 858 ~ Faktor 4 (Zeit)
 - Aufgabe 3: 33T / 2108T ~ Faktor 63 (Größe)
- Ohne Optimierung wären die Zeiten noch viel schlechter

Ändern

- Richtzeiten angeben?
- Waren 10GB zu viel?
- Ergebnisgrößen angeben!
 - Um Probleme bei Aufgabe 2 zu beheben

- Achtung: Die Gruppen-Accounts werden in einer Woche alle gelöscht