TU München, Fakultät für Informatik
Lehrstuhl III: Datenbanksysteme
Prof. Alfons Kemper, Ph.D.

Übung zur Vorlesung *Einsatz und Realisierung von Datenbanken* im SoSe23

Alice Rey, Maximilian Bandle, Michael Jungmair (i3erdb@in.tum.de)
http://db.in.tum.de/teaching/ss23/impldb/

Blatt Nr. 09

Hinweise Für die aktive Teilnahme an der dieswöchigen Übung benötigen Sie Spark auf Ihrem Rechner installiert. Eine Anleitung finden Sie unter `https://db.in.tum.de/teaching/ss23/impldb/Spark_preparation.pdf`.

Hausaufgabe 1

HyPer schafft 120.000 Transaktionen pro Sekunde. Pro Transaktion werden 120 Byte in die Log geschrieben. Berechnen Sie den benötigten Durchsatz zum Schreiben der Log.

Die Datenbank läuft für einen Monat und stürzt dann ab. Es wurde kein Snapshot erstellt. Berechnen Sie die Recoveryzeit. Gehen Sie davon aus, dass die Recovery durch die Festplatte limitiert ist (100 MiB / s). Wieviel Log Einträge werden pro Sekunde recovert?

Hausaufgabe 2

Gegeben eine Tabelle *Produkte* mit folgendem Schema und 10000 Einträgen:

Id (8 Byte) | Name (32 Byte) | Preis (8 Byte) | Anzahl (8 Byte)

Wieviele Daten werden für folgende Queries in die CPU-Caches geladen? Unterscheiden sie jeweils zwischen Row und Column Store.

1. *select * from Produkte*
2. *select Anzahl from Produkte*

Hausaufgabe 3

Sie sollen für die Alexander-Maximilians-Universität (AMU) ein Hauptspeicherdatenbanksystem optimieren. In dem System sind die Daten aller Studenten gespeichert. Schätzen Sie für jede der untenstehenden Anfragen einzeln, ob ein Row- oder Column-Store besser geeignet ist.

Relationen

Studenten: MatrNr (8 Byte), Name (48 Byte), Studiengang (4 Byte), Semester (4 Byte)

MatrNr ist der Primärschlüssel der indiziert ist.

Anfragen:

1. select * from Studenten;
2. select Semester, count(*) from Studenten group by Semester;
3. select Name, Studiengang, Semester from Studenten where MatrNr = 42;
4. select Studiengang from Studenten where MatrNr = 42;
5. select * from Studenten where Semester < 5;

6. select * from Studenten where Semester = 25;
7. insert into studenten values(4242, Max Meyer, Info, 7);

Hausaufgabe 4

Rekonstruieren Sie die ursprüngliche SQL-Anfrage aus dem folgenden (Pseudo-)Code eines codegenerierenden Datenbanksystems. Welche Art von physikalischem Join wurde benutzt? Handelt es sich um Column- oder Row-Store?

```
struct Student { int matrnr; std::string name; int semester; };
struct Hoeren { int matrnr; int vorlnr; };
struct Result { int vorlnr; int a; };

std::vector<Result> compute(std::vector<Student>&ses,std::vector<Hoeren>&hs){
       std::unordered_multimap<int, Hoeren*> h_map;
       std::unordered_map<int, Student*> s_map;
       for (auto &h : hs)
              h_map.insert(std::make_pair(h.matrnr, &h));
       for (auto &s : ses)
              s_map.insert(std::make_pair(s.matrnr, &s));

       std::unordered_map<int, int> count_map;
       std::unordered_map<int, int> sum_map;
       for (auto &h : hs) {
              count_map.insert(std::make_pair(h.vorlnr, 0));
              sum_map.insert(std::make_pair(h.vorlnr, 0));
       }
       for (auto &h : h_map) {
              sum_map[h.second->vorlnr] += s_map[h.first]->semester;
              count_map[h.second->vorlnr]++;
       }
       std::vector<Result> res;
       for (auto &r : sum_map)
              res.push_back({ r.first, r.second / count_map[r.first] });
       return res;
}
```

Hausaufgabe 5

Führen Sie die folgenden Abfragen in der Spark-Shell aus. Als Grundlage für die Abfragen dient das TPC-H Schema. Laden Sie dazu die TPC-H Daten wie in der Vorlesung gezeigt in die Spark-Shell.

(a) Ermitteln Sie pro Marktsegment die Anzahl der Bestellungen in 1997.

(b) Ermitteln Sie die Zahl der Kunden und Lieferanten pro Land.

(c) Ermitteln Sie die Stückzahlen der verschiedenen Bauteile in Deutschland.

(d) Ermitteln Sie, welche Kunden kein *goldenrod lavender spring chocolate lace* bestellt haben.

Hausaufgabe 6

Führen Sie die folgenden Abfragen in der Spark-Shell aus. Als Grundlage für die Abfragen dient das TPC-H Schema. Laden Sie dazu die TPC-H Daten wie in der Vorlesung gezeigt in die Spark-Shell.

(a) Laden Sie die `region.tbl` Datei als DataFrame Objekt in die Spark-Shell.

(b) Ermitteln Sie die Namen aller Regionen.

(c) Ermitteln Sie die Zahl der Länder die nicht in Europa liegen.

(d) Ermitteln Sie die größte Bestellung aus dem Jahr 1996.

(e) Ermitteln Sie welcher europäische Kunde im Jahr 1996 am meisten Geld ausgegeben hat.

(f) Ermitteln Sie welche Unternehmen keine Kunden in Europa haben.