December 15, 2016
Arnab Chakrabarti, Rihan Hai, Sandra Geisler

Implementation of Databases (WS 16/17)

Exercise 5

Due until January 10, 2017, 2pm.
Please submit your solution *in a single PDF file* before the deadline to the L^{2}P system!
Please submit solutions in groups of three students.

Exercise 5.1 (Query Optimization and Cost Estimation) (12 pts)

Consider the following relational schema and SQL query. The schema captures information about employees, departments, and finances (organized on a per department basis).

$$Emp(\underline{eid}, did, sal, hobby)$$
$$Dept(\underline{did}, dname, floor, phone)$$
$$Finance(\underline{did}, budget, sales, expenses)$$

$$Emp[did] \subseteq Dept[did]$$
$$Finance[did] \subseteq Dept[did]$$

```
SELECT d.dname, f.budget
FROM Emp e, Dept d, Finance f
WHERE e.did=d.did AND d.did=f.did AND d.floor=1
    AND e.sal>59000 AND e.hobby='yodeling'
```

1. Identify a relational algebra tree (or a relational algebra expression if you prefer) that reflects the order of operations a decent query optimizer would choose (applying the heuristics selection before join, and projections are done as many and as early as possible).

2. Suppose that the following additional information is available: Unclustered B+-tree indexes exist on Emp.did, Emp.sal, Dept.floor, Dept.did, and Finance.did. Furthermore, assume that the costs for a tree traversal (from root to a leaf) are 3, additional costs for scanning leaves sequentially can be ignored. The system's statistics indicate that employee salaries range from 10.000 to 60.000, employees enjoy 200 different hobbies, and the company owns two floors in the building. All attribute values are uniformly distributed. There are a total of 50.000 employees and 5.000 departments (each with corresponding financial information, i.e., relation Finance has also 5.000 records) in the database. The DBMS used by the company has just one join method available: index nested loops.

(a) For each of the query's base relations (Emp, Dept, and Finance) estimate the number of tuples that would be initially selected from that relation if all of the non-join predicates on that relation were applied to it before any join processing begins.

(b) Given your answer to the preceding question, what would be the total cost for a query plan that first performs the selections on Emp, then joins Emp with Dept, and then the result is joined with Finance. Assume again that selections are done before join, projections are done as many and as early as possible, and that the indexes are used if applicable. You can also assume, that intermediate results do not have to be stored (either, they fit into memory or they are pipelined to the next operator).

Exercise 5.2 (Information Integration) **(6 pts)**

Suppose we have a virtual data integration system with the following mediated schema:

$$\mathcal{G} : Movie(Title, Director, Genre), Schedule(Cinema, Title, Time)$$

There are two local data sources V_1 and V_2, against which we can query data. The first view V_1 has the schema $Times(Title, Time, Director)$. The second view V_2 has the schema $Cinemas(Cinema, Genre, Title)$.

1. Provide a Local-As-View mapping between the two view predicates V_1 and V_2 and $\mathcal{G}$.

2. Now suppose V_1 and V_2 are views defined over a database which consists of the following relations:

   Movies(Title,Director,Year,Genre)
   Director(Director, Age)
   Genre(Title, Genre)
   Playing(Cinema,Title,Time)

   The two views are associated via a Global-As-View mapping to the stored relations in the database.

   $$V_1(Title, Time, Director) \leftarrow Movies(Title, Director) \wedge Playing(Title, Time)$$
   $$V_2(Cinema, Genre, Title) \leftarrow Playing(Cinema, Title) \wedge Genre(Title, Genre)$$

   Please do a rewriting of the following query against the view V_1 such that you query the global schema using the GAV mapping:

```
SELECT Title, Time, Director FROM Times
WHERE Title="The Hobbit: The Battle of Five Armies" AND
      Time="23.12.2014"
```

Exercise 5.3 (Serialization and Recovery) **(12 pts)**

Consider the following schedules:

- $s_1 = w_2(y)w_3(x)r_1(x)c_1r_3(y)c_3w_2(x)c_2$
- $s_2 = w_2(y)w_3(z)r_1(x)c_1w_2(x)c_2r_3(y)c_3$
- $s_3 = w_2(y)w_3(y)r_1(x)c_1w_2(x)r_3(y)c_3c_2$

1. Write down the conflict relations $\text{conf}(s_n)$ of the schedules $s_n, n \in 1, 2, 3$.
2. Determine for each schedule $s_n, n \in \{1, 2, 3\}$ if it belongs to the classes CSR, OCSR and CO. Proof your decision.
3. Determine for each schedule $s_n, n \in \{1, 2, 3\}$, if belongs to the recovery classes RC, ACA and ST.
   Proof your decision.