Verteilte und föderierte Datenbanken
Dr. Eike Schallehn / Ingolf Geist

Exercise 8 - Distributed Query Processing II

based on [1]

The following relation schema is given:

```
EMPLOYEE(ENR, ENAME, JOB, SALARY)
PROJECT(PNR, ENAME, BUDGET)
ASSIGNMENT(ENR, PNR, DURATION)
```

1. Data Localization (hybrid Fragmentation)

The relation EMPLOYEE is fragmented as follows:

$$\texttt{EMPLOYEE}_1 = \pi_{\texttt{ENR},\texttt{ENAME}}(\sigma_{\texttt{ENR}<20.000}(\texttt{EMPLOYEE}))$$
$$\texttt{EMPLOYEE}_2 = \pi_{\texttt{ENR},\texttt{JOB},\texttt{SALARY}}(\sigma_{\texttt{ENR}<20.000}(\texttt{EMPLOYEE}))$$
$$\texttt{EMPLOYEE}_3 = \sigma_{\texttt{ENR}\geq 20.000}(\texttt{EMPLOYEE})$$

What is the initial fragment expression for the following query:

```
SELECT ENAME FROM EMPLOYEE WHERE ENR=4711
```

Perform algebraic optimization!

2. Simple Join-Strategies

Given $card(R) = 10.000$, $card(S) = 1.000$, $JSF(R \bowtie S) = 0,001$ for 2 relations R and S. Each relation has 5 attributes. Which communication costs result for Ship Whole (SW) and Fetch as needed (FAN) strategies for join processing on nodes at N_R or N_S?

3. Ship-Whole vs. Semi-Join vs. Bit Vector-Join

The following query on EMPLOYEE and ASSIGNMENT has to be processed:

```
SELECT E.ENR, ENAME, JOB, PNR, DURATION
FROM EMPLOYEE E, ASSIGNMENT A
WHERE E.ENR=A.ENR AND E.SALARY>60.000
```

Furthermore, the following statistics are available: $card(\texttt{EMPLOYEE}) = 1.000$, $card(\texttt{ASSIGNMENT}) = 1.500$; both relations are stored on different nodes. The query is initiated on a third node N and the result must be available there. The salary condition is satisfied by 20% of the employees ($SF = 0,2$); 25% of the employees do not work for any specific project.

Evaluate the join processing strategies (#Messages, #Values):

- Ship-Whole; join processing on node $N_{\texttt{ASSIGNMENT}}$
- Ship-Whole; join processing on node N

- Semi-Join; join processing on node $N_{\texttt{EMPLOYEE}}$
- Semi-Join; join processing on node N
- Bit Vector-Join; join processing on node N

Before join processing all executable selections and projections should be performed. The length of the bit vector should be equivalent to 5 data values. Using the hash filtering increases the size of the intermediate result by 5%.

4. Multi-Way Joins

Estimate the the communication costs for the following query

```
SELECT * FROM EMPLOYEE E, PROJECT P, ASSIGNMENT A
WHERE E.ENR=A.ENR AND P.PNR=A.PNR AND JOB='SW-Developer'
```

using the Ship-Whole- and Semi-Join-strategy. Each of the three relations is stored on a different node. Furthermore, the following statistics are known: $card(\texttt{EMPLOYEE}) = 1.000$, $card(\texttt{ASSIGNMENT}) = 1.500$, $card(\texttt{PROJECT}) = 200$. The query is initiated at node $N_{\texttt{EMPLOYEE}}$ and the result must be returned there. The job selection is satisfied by 10% of the employees ($SF = 0,1$); 25% of the employees work in no specific project.

Literatur

[1] Erhard Rahm. Mehrrechner-Datenbanksysteme: Grundlagen der verteilten und parallelen Datenbankverarbeitung. Addison-Wesley Bonn, 1994