The following relation schema is given:

\[
\begin{align*}
\text{EMPLOYEE} & : (ENR, ENAME, JOB, SALARY) \\
\text{PROJECT} & : (PNR, ENAME, BUDGET) \\
\text{ASSIGNMENT} & : (ENR, PNR, DURATION)
\end{align*}
\]

1. Data Localization (hybrid Fragmentation)

The relation \text{EMPLOYEE} is fragmented as follows:

\[
\begin{align*}
\text{EMPLOYEE}_1 & = \pi_{ENR, ENAME} (\sigma_{ENR < 20.000} (\text{EMPLOYEE})) \\
\text{EMPLOYEE}_2 & = \pi_{ENR, JOB, SALARY} (\sigma_{ENR < 20.000} (\text{EMPLOYEE})) \\
\text{EMPLOYEE}_3 & = \sigma_{ENR \geq 20.000} (\text{EMPLOYEE})
\end{align*}
\]

What is the initial fragment expression for the following query:

\[
\text{SELECT ENAME FROM EMPLOYEE WHERE ENR=4711}
\]

Perform algebraic optimization!

2. Simple Join-Strategies

Given \(\text{card}(R) = 10,000\), \(\text{card}(S) = 1,000\), \(JSF(R \bowtie S) = 0.001\) 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 \(N_S\)?


The following query on \text{EMPLOYEE} and \text{ASSIGNMENT} has to be processed:

\[
\begin{align*}
\text{SELECT E.ENR, ENAME, JOB, PNR, DURATION} \\
\text{FROM EMPLOYEE E, ASSIGNMENT A} \\
\text{WHERE E.ENR=A.ENR AND E.SALARY>60.000}
\end{align*}
\]

Furthermore, the following statistics are available: \(\text{card(EMPLOYEE)} = 1,000\), \(\text{card(ASSIGNMENT)} = 1,500\); 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 \(N_{\text{ASSIGNMENT}}\)
- Ship-Whole; join processing on node \(N\)
• Semi-Join; join processing on node $N_{\text{EMPLOYEE}}$
• 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: $\text{card}(\text{EMPLOYEE}) = 1.000$, $\text{card}(\text{ASSIGNMENT}) = 1.500$, $\text{card}(\text{PROJECT}) = 200$. The query is initiated at node $N_{\text{EMPLOYEE}}$ 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