$T(W)=10^3$, $T(R)=10^6$, $T(S)=10^5$, $V(R.C)=10^2$

T( ): total # of tuples | π: attribute cond. | γ group by
V( ): unique values | σ: tuple

New Problem Consider an alternate optimizer that produces all possible algebra expressions, including ones where the selection $\sigma_{R.C=1}$ is not applied directly on R. This alternate optimizer also chooses as "optimal" the expression with the smallest sum of intermediate result sizes.

Is the optimal expression still the one we had found above (the second plan)? Justify your answer.

Goal: get under 10^4+10

Plan #1 $\sigma_{R.C=1}((R \bowtie W) \bowtie S)$ $\quad T(a) = T(R \times W) \Rightarrow$

$$\frac{T(R)T(W)}{\max(V(R,B)\,V(W,B))} = 10^3$$

$$T(T(a) \bowtie S) = \frac{T(a)T(S)}{V(S,A)} = \frac{10^3\,10^8}{10^8} = 10^3$$

$\boxed{2\times 10^3 \text{ This plan is better}}$

Pl $(\sigma_{R.C=1}(R \bowtie W)) \bowtie S$ $\quad T(a) = T(R\times W) = T(W) = 10^3$

$$T(\sigma(R\times W)) = \frac{T(a)}{V(R,C)} = \frac{10^3}{10^2} = 10$$

$\boxed{10^3+10}$

This plan is better than original answer

Previous plan is best. The [illegible]

2 Algebra and Estimation

Produce an optimal algebraic expression for the following query over tables R(A, B) and S(A, C, D), where "optimal" means that it has the smallest total size of intermediate results, among all possible algebraic expressions that are equivalent to this query. Write the sizes of all intermediate results.

```
SELECT A, C, AGG(B) AS N
FROM R, S
WHERE S.A = R.A AND S.D = 1
GROUP BY A, C
```

given the following statistics

$$T(R) = 10^9$$
$$V(R, A) = 10^6$$
$$V(R, B) = 10^9$$
$$T(S) = 10^{10}$$
$$V(S, A) = 10^7$$
$$V(S, C) = 10^2$$
$$V(S, D) = 10$$

Assume (the common assumption) that

$$V(R, A) < V(S, A) \Rightarrow \pi_A R \subset \pi_A S$$

a:) $R \bowtie S = \sigma_{D=1} A \Rightarrow T(R \bowtie S) = \min\{T(R), T(S)\} = \{10^9, 10^{10}\} = 10^9$

$T(\sigma_{D=1} A) = \frac{T(A)}{T(A,D)} = \frac{10^9}{10^1} = 108$

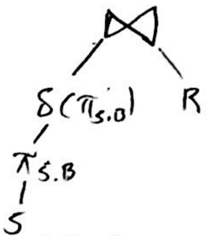
$\boxed{total = 10^9 + 108}$ *

ii) $\sigma_{D=1} S \bowtie D \Rightarrow T(\sigma_{D=1} S) = 10^{10}/10^1 = 10^9$

$T(\sigma_{D=1} S \bowtie R) = \min\{10^9, 10^9\} = 10^9$

total $= 10^9 + 10^9$

$R \ltimes_c S : \pi_{A,B}(R \bowtie_B \delta(\pi_{S.B}(S)))$



$$R \ltimes_c S = \pi_{\bar{A}}(R \bowtie_c S)$$

Except that it is wrong. Provide a correct one.

Equivalences Now, assume relations R, S and T. The notation $c(A)$ refers to a condition that refers to the list of attributes A only. Declare true or false each of the following. If the answer is "no", also provide counterexample.

1. $\sigma_{c(A)}(R \ltimes S) = (\sigma_{c(A)} R) \ltimes S$, where R has a list of attributes A and S has no attributes of A. $\boxed{\text{True}}$

2. $\delta(R \ltimes S) = (\delta R) \ltimes S$ $\boxed{\text{True}}$

3. $(R \triangleright\!< S) \triangleright\!< T = (R \triangleright\!< T) \triangleright\!< S$, where all of R, S and T have a single common attribute A and no pair of R, S and T has a common attribute other than A. $\boxed{\text{True.}}$

Recall, δ is the duplicate elimination operator.

$\sigma_{c(A)}(R \propto S)$

1a)

R

A	B
1	2
1	2
3	4
3	2

S

B	C
2	3
2	3
2	3

$R \propto S \rightarrow \theta_{c(A)}$

A	B
1	2
1	2
3	2

A	B
1	2
1	2

1b) $R\ \sigma_{c(A)=1} \Rightarrow (\sigma_{c(A)=1} R) \propto S$

A	B
1	2
1	2

A	B
1	2
1	2

2) $\delta(R \propto S) = (\delta R) \propto S$

$(R \propto S)$

A	B
1	2
1	2
3	2

A	B
1	2
3	2.

$\delta R \rightarrow (\delta R \propto S)$

A	B
1	2
3	4
3	2
4	null

A	B
1	2
3	2

3. $(R \propto S) \propto T$

* random note, can't match null

R

A	B
1	2
1	2
3	4
3	2
4	null

S

B	C
2	3
2	3
2	3
null	4

T

A	C
1	3
2	2
2	4

a) $(R \propto S) \propto T$

A	B
1	2
1	2
3	2

A	B
1	2
1	2

b) $(R \propto T) \propto S$

A	B
1	2
1	2

A	B
1	2
1	2

int : 4 bytes
float = 8 bytes
32 GB RAM.
$V(D_1, R) = V(D_2, R) = V(D_3, R) = 10^6$
$V([D_1, D_2, D_3], R) = 10^8$
$R(D_1, D_2, D_3, M)$

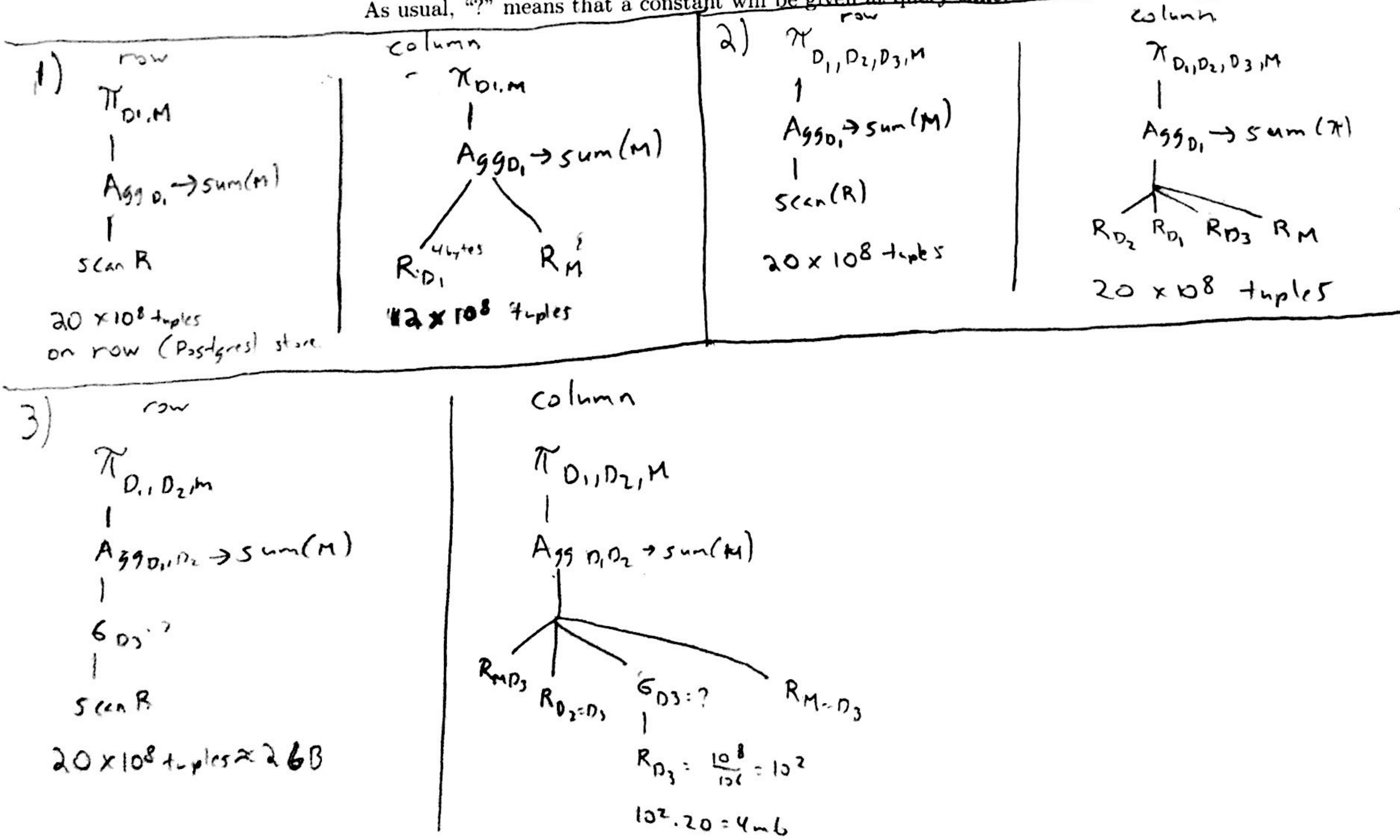
4 Column Databases

Consider the table R(D1,D2,D3,M). You already have a PostGres database and you wonder whether it is worthy to buy a column database in order to accomodate queries on R. Of course, the answer depends on knowing the queries that will be issued. For each of the following queries, declare whether a column database will be significantly better or whether PostGres is good enough(or even better). Just place a circle around the relevant system. "Significantly" means at least close to a multiple, say 2x. A 10% improvement is not significant.

When you consider column databases, assume they do not have indices. (This is not exactly accurate, as column databases also have indices, but we adopt it for the sake of the exercise.) Furthermore, assume that the table is in the order of terabytes and resides in hard disk. The main memory is only 32GB. Each one of the D attributes is an integer and M is a float. Assume that $V(\texttt{D1}, \texttt{R}) = V(\texttt{D2}, \texttt{R}) = V(\texttt{D3}, \texttt{R}) = 10^6$. Assume $V([\texttt{D1}, \texttt{D2}, \texttt{D3}], \texttt{R}) = 10^8$.

1. SELECT D1, SUM(M) FROM R GROUP BY D1: [✓ Column] - Postgres better (≈40%)
2. SELECT D1, D2, D3, SUM(M) FROM R: Column - [✓ Postgres] good enough
3. SELECT D1, D2 SUM(M) FROM R WHERE D3=?: [✓ Column] - Postgres significantly better ≈4mb << 2GB

As usual, "?" means that a constant will be given at query time.



1)
row
$\pi_{D1,M}$
$Agg_{D_1} \to sum(M)$
scan R
20 × 10^8 tuples
on row (Postgres) store
column
$\pi_{D1,M}$
$Agg_{D_1} \to sum(M)$
R_{D_1} 4bytes
R_M 8
12 × 10^8 tuples
2)
row
$\pi_{D_1,D_2,D_3,M}$
$Agg_{D_1} \to sum(M)$
scan(R)
20 × 10^8 tuples
column
$\pi_{D_1,D_2,D_3,M}$
$Agg_{D_1} \to sum(M)$
R_{D_2} R_{D_1} R_{D_3} R_M
20 × 10^8 tuples
3)
row
$\pi_{D_1,D_2,M}$
$Agg_{D_1,D_2} \to sum(M)$
$\sigma_{D_3 = ?}$
scan R
20 × 10^8 tuples ≈ 2GB
column
$\pi_{D_1,D_2,M}$
$Agg_{D_1,D_2} \to sum(M)$
R_{MD_3} $R_{D_2 = D_3}$ $\sigma_{D3 : ?}$ $R_{M - D_3}$
$R_{D_3} = \frac{10^8}{10^6} = 10^2$
$10^2 \cdot 20 = 4mb$