Final Exam

DSE201, Winter 2016

Name: Ryan Riopelle

Brief Directions:

- Write clearly!
- Good luck!

DSE 201 - Final - Ryan Riopelle

1.)

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

B is a key of W. B in non-null FK that references R.B.

A is Key of S and R.A is non null FK that references S.A.

Plans

1) $\sigma_{R.C=1}((R \bowtie S) \bowtie W)$
2) $\sigma_{R.C=1}((R \bowtie W) \bowtie S)$
3) $(\sigma_{R.C=1}(R \times S)) \bowtie W$
4) $(\sigma_{R.C=1}(R \bowtie W)) \bowtie S$

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

b) $= T(T(a) \bowtie S) = \frac{T(a)T(S)}{V(S,A)} = \frac{(10^3)(10^5)}{10^8} = 10^3$

Intermediat = 2×10^3

a) $T(a) = T(R \bowtie W) = T(W) = 10^3$

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

Intermediate = $10^3 + 10$

Plan 2 and 4 are less costly

Name: Ryan Riopelle

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

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

(1)

$\gamma_{A,C;\ AGG(B)\to N}(R \bowtie (\sigma_{S.D=1} S))$

a) $T(\sigma_{S.D=1} S) = \frac{T(S)}{V(S,D)} = \frac{10^{10}}{10} = 10^9$

b) $T(R \bowtie T(a)) = \frac{T(R)T(a)}{\max(V(R,A), V(S,A))} = \frac{10^9 \cdot 10^9}{10^7} = 10^{11}$

(10^6, 10^7)

Intermediate: $10^9 + 10^{11}$

(2)

$\gamma_{A,C;\ AGG(B)\to N}(\sigma_{S.D=1}(R \bowtie S))$

a) $T(a) = T(R \bowtie S) = \frac{T(R)\,T(S)}{\max(V(R,A), V(S,A))} = \frac{10^9 \cdot 10^{10}}{10^7} = 10^{12}$

b) $T(\sigma_{S.D=1} T(a)) = \frac{T(a)}{V(S,D)} = \frac{10^{12}}{10} = 10^{11}$

Intermediate results: $10^{12} + 10^{11}$

Plan 1 is better!

Name: Ryan Riopelle

3)

$$R \overline{\ltimes}_c S = \pi_{\bar{A}}(R \bowtie_c S)$$

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. $\sigma_{c(A)}(R \overline{\ltimes} S) = (\sigma_{c(A)}R) \overline{\ltimes} S$, where R has a list of attributes A and S has no attributes of A. — TRUE

2. $\delta(R \overline{\ltimes} S) = (\delta R) \overline{\ltimes} S$ — True

3. $(R \overline{\rtimes} S) \overline{\rtimes} T = (R \overline{\rtimes} T) \overline{\rtimes} S$, 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. — TRUE

Recall, δ is the duplicate elimination operator.

Queries

$C \overline{\ltimes}_{cid = e.class} e$

Algebraic Definition

$R \overline{\ltimes}_c S = \pi_A (R \overline{\ltimes}_c \delta_c S)$

Name: Ryan Riopelle

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 $V(\texttt{[D1,D2,D3]},\texttt{R}) = 10^8$.

1. `SELECT D1, SUM(M) FROM R GROUP BY D1`: Column - Postgres
2. `SELECT D1, D2, D3, SUM(M) FROM R`: Column - Postgres or equal
3. `SELECT D1, D2 SUM(M) FROM R WHERE D3=?`: Column - Postgres

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

- See work for more Info.

Name: Ryan Riopelle

Database Final - Problem 4

4-1) Table R(D1, D2, D3, M), Size = Tera byte
$V(D1, R) = V(D2, R) = V(D3, R) = 10^6$
$V([D1, D2, D3], R) = 10^8$

Postgres: $O(\sigma_{D1, D2, D3, D4, m} R) = 10^8 \times [4+4+4+8] = 20 \times 10^2$ byts
$= 1.86$ GB

Column: $O(\sigma_{D3, m} R) = 10^2 \times [4+8] = 12 \times 10^8$ bytes $= 1.12$ GB

Column DB = 40% faster

4-2) Select D1, D2, D3, Sum(M) From R

About equal, Postgres

4-3) Select D1, D2 Sum(M) From R where
D3 = ?
Postgre: 1.86 GB
Column

$$T(\sigma_{D3=R}) = \frac{T(R)}{V(D3, R)} = \frac{V([D1, D2, D3], R)}{V(D3, R)}$$

$$O(\sigma_{D1, D2 m} R) = (4+4+8) \times 10^2 = 16 \times 10^2$$

$$O(\sigma_{D3=R}) = 4 \times 10^8 = 381 \text{ Mb}$$
$$O(\text{Total}) = (4 \times 10^8) + ((6 \times 10^2) \simeq 0.381 \text{ GB}$$

$\frac{1.86 - .381}{1.86} =$ **80% faster - column**