Advanced Data Engineering: Assignment 7

NGUYEN T. Hoang - SID: 15M54097
(ホアン)

Fall 2015, W831 Tue. Period 5-6
Due date: 2015/12/08

Problem

- Consider conditions for making disk-I/O costs of the GRACE hash join and Hybrid hash join superior to those of simple hash join, respectively (GRACE vs. Simple, Hybrid vs. Simple).
- QUIZ: Consider an SQL sentence to derive the result given in the lecture note.

Answer

Based on the cost estimation model and the *Comparision of Hash Join Algorithms* table, we have the following notations:

$$\text{Let: } \alpha = \frac{|M|}{|R|}. \text{ I will make conditions based on } \alpha. \tag{1}$$

The cost estimation of Simple Hash Join is re-written using α as follow:

- Number of disk I/O:

$$SH_{I/Os} = \frac{|R|}{|M|} \times (|R| + |S|) = \frac{1}{\alpha} \times (|R| + |S|)$$

- The cost of hashing:

$$SH_{Hash\_Cost} = \frac{|R| + |M|}{2|M|}(\{R\} + \{S\}) = \frac{\alpha + 1}{2\alpha}(\{R\} + \{S\})$$

1 GRACE Hash Join

The cost estimation of GRACE Hash Join given in the lecture note is:

- Number of disk I/O required:

$$GH_{I/Os} = 3 \times (|R| + |S|)$$

- The cost of hashing:

$$GH_{Hash\_Cost} = 2 \times (\{R\} + \{S\})$$

Disk I/O ratio between GRACE Hash Join and Simple Hash Join is:

$$\mu_{GRACE-Simple} = \frac{GH_{I/Os}}{SH_{I/Os}} = 3 \times \alpha$$

As we can see here, the disk I/O ratio between GRACE Hash Join and Simple Hash Join is directly proportional to our parameter α. Intuitively this means that if we have a small memory and a large relation table (small α), then the number of disk I/O performs by GRACE Hash Join will be much smaller than that of Simple Hash Join. We have:

- $\alpha \leq \frac{1}{3}$: GRACE Hash Join has equal or worse disk I/O cost than Simple Hash Join (larger number of Disk I/O). In this case, at least a third of the relational table can fit into memory.
- $\alpha < \frac{1}{3}$: GRACE Hash Join has better disk I/O cost than Simple Hash Join (smaller number of Disk I/O). In this case, the size of the relational table is at least 3 times bigger than the memory.
- $\alpha \ll \frac{1}{3}$: GRACE Hash Join is superior in term of I/O cost compare to Simple Hash Join. This is the practical case, where we have limitted amount of memory, but the size of the relational table is very big.

2 Hybrid Hash Join

The cost estimation of Hybrid Hash Join given in the lecture note is rewritten as follow:

- Number of disk I/O required:

$$HH_{I/Os} = (3 - 2\frac{|M|}{|R|}) \times (|R| + |S|) = (3 - 2\alpha) \times (|R| + |S|)$$

- The cost of hashing:

$$HH_{Hash\_Cost} = 2 \times (\{R\} + \{S\})$$

Disk I/O ratio between Hybrid Hash Join and Simple Hash Join is:

$$\mu_{Hybrid-Simple} = \frac{HH_{I/Os}}{SH_{I/Os}} = 3\alpha - 2\alpha^2$$

This model is designed for practical settings, therefore we assume that $0 < \alpha < \frac{3}{2}$. The Disk I/O cost estimation of Hybrid Hash Join is stated as follow:

- $0 < \alpha < \frac{1}{2}$: Hybrid Hash Join has better disk I/O cost than Simple Hash Join (smaller number of Disk I/O).
- $\frac{1}{2} \leq \alpha \leq 1$: Hybrid Hash Join has equal or worse disk I/O cost than Simple Hash Join. The worst case is when $\alpha = 3/4$.
- $1 < \alpha < \frac{3}{2}$: Hybrid Hash Join has better disk I/O cost than Simple Hash Join.

3 QUIZ: SQL sentence

The SQL sentence to derive each saleman in R and his/her sale count of some product ID in the product table S is:

```
1 SELECT Salesman, COUNT(*) FROM R
2 WHERE ProductID IN (
3     SELECT ProductID FROM S)
4 GROUP BY Salesman;
```