Advanced Data Engineering: Assign. 12

NGUYEN T. Hoang - SID: 15M54097
(ホアン)

Fall 2015, W831 Tue. Period 5-6
Due date: 2016/01/19

Problem

- Roughly estimate the execution time for the 4 distributed algorithms with the following assumptions:
  - Cardinality of the relation R = 1,000,000; S = 500,000.
  - Total length of a tuple: $st_R = 1,000B$; $st_S = 2,000B$.
  - Disk transfer bandwidth: $B_{disk} = 10MB/s$.
  - Network bandwidth: $B_{net} = 5MB/s$.
  - Selectivity: $\alpha = 10\%$, $\beta = 10\%$, $\gamma = 10\%$,
  - Hash Bit Vector: Use 1 bit for each tuple.

Question: Cost estimation for distributed algorithms.

Estimate cost for Naïve, Semi-Join, 2-Way Semi-Join and Hashed Bit Vector distributed join algorithms.

Answer:

Naïve According to the lecture note, the cost estimation for naïve algorithm is:

$$\mathcal{C}_{\text{Naïve}} \approx \mathcal{C}_D(5S + 3R) + \mathcal{C}_C(S)$$

Since the data size of the relations are same, I assume relation S is sent. Replace given information about data, we have:

$$\begin{aligned}
\mathcal{C}_{\text{Naïve}} &\approx \mathcal{C}_D(5S + 3R) + \mathcal{C}_C(S) \\
&= \frac{5 \times \{S\} \times st_S + 3 \times \{R\} \times st_R}{B_{disk}} + \frac{\{S\} \times st_S}{B_{net}} \\
&= \frac{5 \times 500,000 \times 2,000 + 3 \times 1,000,000 \times 1,000}{10,000,000} + \frac{500,000 \times 2,000}{5,000,000} \\
&= 800 + 1000 = 1800 \text{ seconds} = 30 \text{ minutes.}
\end{aligned}$$

Semi-Join According to the lecture note, the cost estimation for semi-join algorithm when we assume that projection can be overlapped on I/O is:

$$\mathcal{C}_{\text{SJ}} \approx \mathcal{C}_D(4S + 5\alpha S + 3R + 5\beta R) + \mathcal{C}_C(\alpha S + \beta R)$$

Replace given information about data, we have:

$$\begin{aligned}\mathcal{C}_{\text{SJ}} &\approx \mathcal{C}_D(4S + 5\alpha S + 3R + 5\beta R) + \mathcal{C}_C(\alpha S + \beta R)\\ &= \frac{(4+5\alpha)\times 500,000 \times 2,000 + (3+5\beta)\times 1,000,000 \times 1,000}{10,000,000}\\ &\quad + \frac{\alpha \times 500,000 \times 2,000 + \beta \times 1,000,000 \times 1,000}{5,000,000}\\ &= 800 + 40 = 840 \text{ seconds} = 14 \text{ minutes.}\end{aligned}$$

2-Way Semi-Join According to the lecture note, the cost estimation for 2-way semi-join algorithm is:

$$\mathcal{C}_{\text{2WSI}} \approx \mathcal{C}_D(4R + 5\beta R + 3\alpha S + 2\beta\gamma R + 2\alpha\gamma S) + \mathcal{C}_C(\alpha\gamma S + \beta R)$$

Replace given information about data, we have:

$$\begin{aligned}\mathcal{C}_{\text{2WSI}} &\approx \mathcal{C}_D(4R + 5\beta R + 3\alpha S + 2\beta\gamma R + 2\alpha\gamma S) + \mathcal{C}_C(\alpha\gamma S + \beta R)\\ &= \frac{(2\alpha\gamma+3\alpha)\times 500,000 \times 2,000 + (4+5\beta+2\beta\gamma)\times 1,000,000 \times 1,000}{10,000,000}\\ &\quad + \frac{\alpha\gamma \times 500,000 \times 2,000 + \beta \times 1,000,000 \times 1,000}{5,000,000}\\ &= 484 + 22 = 506 \text{ seconds} = 8.4 \text{ minutes.}\end{aligned}$$

Hashed Bit Vector According to the lecture note, the cost estimation for hashed bit vector algorithm:

$$\mathcal{C}_{\text{HB}} \approx \mathcal{C}_D(2S + 5\alpha S + 4R + 4\beta R) + \mathcal{C}_C(H_x + H_y + \alpha S)$$

Replace given information about data, we have:

$$\begin{aligned}\mathcal{C}_{\text{HB}} &\approx \mathcal{C}_D(2S + 5\alpha S + 4R + 4\beta R) + \mathcal{C}_C(H_x + H_y + \alpha S)\\ &= \frac{(2+5\alpha)\times 500,000 \times 2,000 + (4+4\beta)\times 1,000,000 \times 1,000}{10,000,000}\\ &\quad + \frac{\alpha \times 500,000 \times 2,000 + 1,000,000 + 500,000}{5,000,000}\\ &= 690 + 20.3 = 710.3 \text{ seconds} = 11.8 \text{ minutes.}\end{aligned}$$