In-Situ Data Processing

Weijie Zhao

04/26/2024

Raw Data Format

- Twitter Data
- Sloan Digital Sky Survey
  - photoPrimary   509 attributes
  - Only 74 attributes are referenced in 70% queries

- What data to load to maximize the query performance?

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
Q_1	X	X						
Q_2	X	X	X	X				
Q_3			X	X	X			
Q_4		X		X		X		
Q_5	X		X	X	X		X	
Q_6	X	X	X	X	X	X	X	

Parameter	Description
$\|R\|$	number of tuples in relation R
S_{RAW}	size of raw file
$SPF_j, j = \overline{1,n}$	size of attribute j in processing format
B	size of storage in processing format
$band_{IO}$	storage bandwidth
$T_{t_j}, j = \overline{1,n}$	time to tokenize an instance of attribute j
$T_{p_j}, j = \overline{1,n}$	time to parse an instance of attribute j
$w_i, i = \overline{1,m}$	weight for query i

$$\text{minimize} \quad T_{load} + \sum_{i=1} w_i \cdot T_i \quad \text{subject to constraints:}$$

$$C_1: \sum_{j=1}^{n} save_j \cdot SPF_j \cdot |R| \leq B$$

$$C_2: read_{ij} \leq save_j;\ i = \overline{1,m},\ j = \overline{1,n}$$

$$C_3: save_j \leq p_{0j} \leq t_{0j} \leq raw_0;\ j = \overline{1,n}$$

$$C_4: p_{ij} \leq t_{ij} \leq raw_i;\ i = \overline{1,m},\ j = \overline{1,n}$$

$$C_5: t_{ij} \leq t_{ik};\ i = \overline{0,m},\ j > k = \overline{1,n-1}$$

$$C_6: read_{ij} + p_{ij} = 1;\ i = \overline{1,m},\ j = \overline{1,n},\ A_j \in Q_i$$

$$T_{load} = raw_0 \cdot \frac{S_{RAW}}{band_{IO}} + |R| \cdot \sum_{j=1}^{n} \left(t_{0j} \cdot T_{t_j} + p_{0j} \cdot T_{p_j} + save_j \cdot \frac{SPF_j}{band_{IO}} \right)$$

$$T_i = raw_i \cdot \frac{S_{RAW}}{band_{IO}} + |R| \cdot \sum_{j=1}^{n} \left(t_{ij} \cdot T_{t_j} + p_{ij} \cdot T_{p_j} + read_{ij} \cdot \frac{SPF_j}{band_{IO}} \right)$$

Definition 1 (k-element cover). *Given a set of n elements $R = \{A_1, \ldots, A_n\}$, m subsets $W = \{Q_1, \ldots, Q_m\}$ of R, such that $\bigcup_{i=1}^{m} Q_i = R$, and a value k, the objective in the k-element cover problem is to find a size k subset R' of R that covers the largest number of subsets Q_i, i.e., $Q_i \subseteq R'$, $1 \leq i \leq m$.*

Definition 2 (minimum k-set coverage). *Given a set of n elements $R = \{A_1, \ldots, A_n\}$, m subsets $W = \{Q_1, \ldots, Q_m\}$ of R, such that $\bigcup_{i=1}^{m} Q_i = R$, and a value k, the objective in the minimum k-set coverage problem is to choose k sets $\{Q_{i_1}, \ldots Q_{i_k}\}$ from W whose union has the smallest cardinality, i.e., $\left|\bigcup_{j=1}^{k} Q_{i_j}\right|$.*

Algorithm 1 Reduce k-element cover to minimum k'-set coverage

Input: Set $R = \{A_1, \ldots, A_n\}$ and m subsets $W = \{Q_1, \ldots, Q_m\}$ of R; number k' of sets Q_i to choose in minimum set coverage

Output: Minimum number k of elements from R covered by choosing k' subsets from W

1: **for** $i = 1$ to n **do**
2:     *res* = ***k-element cover***(W, i)
3:     **if** *res* $\geq k'$ **then return** i
4: **end for**