1 Cost Model Formulas

1.1 Time Cost

The total execution time of all queries is given by:

$$\text{time}_{DB} = \sum_{q=1}^{Q} \left(\frac{\text{vol}_{network}(q)}{\text{bandwidth}_{network}} + \frac{\text{vol}_{RAM}(q)}{\text{bandwidth}_{RAM}} \right) \times freq(q)$$

1.1.1 Communication Volume

Filter queries

$$\text{vol}_{network}(q) = S \cdot size_{query} + res_q \cdot size_{msg}$$

Aggregate queries

$$\text{vol}_{network}(q) = S \cdot size_{query} + shuffle \cdot size_{msg} + res_q \cdot size_{msg}$$

1.1.2 RAM Volume

$$\text{vol}_{RAM}(q, n) = index_q + sel_{att} \cdot coll_{q,n} \cdot size_{doc}(q)$$

The global RAM volume processed is:

$$\text{vol}_{RAM}(q) = \max\left(\text{vol}_{RAM}(q, 1), \ldots, \text{vol}_{RAM}(q, n)\right)$$

1.2 Environmental Cost

1.2.1 Network Carbon Impact

$$impact_{network}(q) = \text{vol}_{network}(q) \times CO2_{network}$$

1.2.2 RAM/CPU Carbon Impact

$$impact_{RAM}(q) = \text{vol}_{RAM}(q) \times CO2_{RAM}$$

1.2.3 Total Environmental Impact

$$impact(DB) = \sum_{q=1}^{Q} \left(impact_{network}(q) + impact_{RAM}(q)\right) \times freq(q)$$

1.3 Financial Cost

1.3.1 Monthly Price

$$price_{DB} = price_s \cdot \max\left(\frac{vol_{DB} \cdot 3}{capacity_{storage}}, \frac{vol_{DB}}{capacity_{RAM}} \cdot 2\right) + externalFees \cdot \sum_{q=1}^{Q} vol_{external}(q) \cdot freq(q)$$

2 Sharding, Indexes, and Algorithms

2.1 When to Use Sharding

- Use sharding when the dataset does not fit on a single server.
- Use sharding when the main queries filter on the sharding key.
- Avoid sharding when most queries filter on non-shard attributes.

2.1.1 Data Access Cost

$$S = \begin{cases} 1 & \text{if filtering on the sharding key} \\ \#shards & \text{otherwise} \end{cases}$$

2.2 When to Use Indexes

- Use an index when selectivity is medium or high.
- Use an index for frequent reads on the attribute.
- Avoid indexes when selectivity is low or for heavy write workloads.

2.3 Choice of Algorithm

2.3.1 Index Lookup

Use when filtering on an indexed attribute:

$$\text{vol}_{RAM}(q) = index_q + sel_{att} \cdot coll \cdot size_{doc}$$

2.3.2 Full Scan

Use when no index exists or selectivity is high:

$$sel_{att} = 1$$

2.3.3 Nested Loop Join

(Only for completeness; NoSQL systems avoid joins.)

$$cost = (S + C1') + C1' \times (S + C2')$$