Network Cost Estimation Formulas

Estimated costs of operations, assuming P machines:

- selection   none
- projection   none
- aggregation of R $B(R) / P$
- join of R and S
  - shuffle R and S $B(R) / P + B(S) / P$
  - shuffle only R $B(R) / P$
  - shuffle only S $B(S) / P$
  - broadcast R $B(R)$
  - broadcast S $B(S)$
- union of Q_1 and Q_2 cost of $Q_1 + Q_2$
- difference of Q_1 and Q_2 $B(Q_1) / P + B(Q_2) / P$

Note that these formulas ignore the cost of final output to the client, which is not relevant to cost-based optimization.

Estimated selectivity of conditions:

- For $A = c$, the selectivity is $1 / (\text{\# distinct values of A})$
- For $A < c$, the selectivity is $(c - \text{lowest value of A}) / (\text{highest - lowest value of A})$
- For $c < A < d$, the selectivity is $(d - c) / (\text{highest - lowest value of A})$