Index-Based Join Size Estimation Using Adaptive Sampling

Sergiu Pocol advised by Yogi Joshi

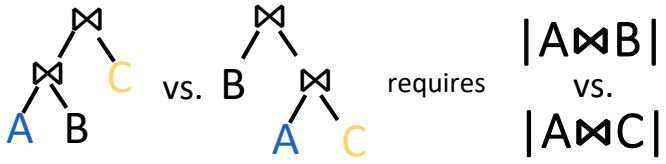
University of Waterloo

SAP

Xi'an, Shaanxi, China
SIGMOD/PODS 2021

Motivation

- Cost-based query optimizers rely on cardinality estimates to avoid a poor query execution plan (QEP)



- *Estimation accuracy was found to be a more significant factor in influencing QEP quality than the sophistication of the cost model [1].*
- Using uniformity and independence assumptions for filtered join size estimation can fail on skewed and correlated data
- Such errors can propagate larger ones in higher-order joins, so improving estimation quality in a "bottom-up" fashion can greatly improve query performance [2]

[1] Leis, V., Gubichev, A., Mirchev, A., Boncz, P., Kemper, A., & Neumann, T. (2015). How good are query optimizers, really? *Proceedings of the VLDB Endowment*, *9*(3), 204–215. https://doi.org/10.14778/2850583.2850594

[2] Leis, V., Radke, B., Gubichev, A., Kemper, A., & Neumann, T. (2017). Cardinality estimation done right: Index-based join sampling. *CIDR*.

Problem

Address the need for uniformity or independence assumptions in two-table join size estimation in SAP IQ

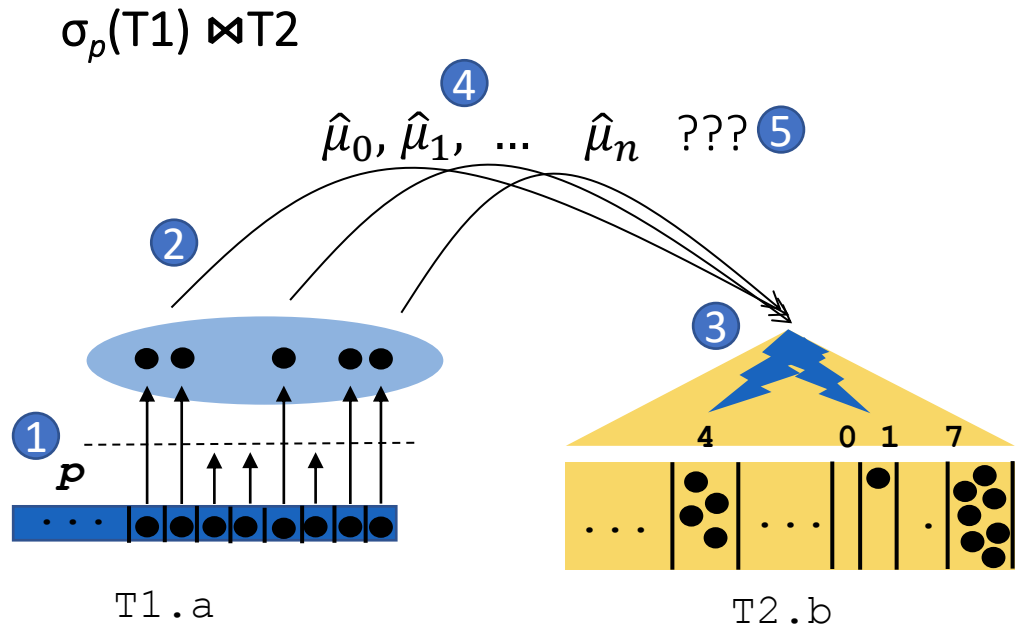
- Minimal overhead
- 'Humble' estimator that defaults to traditional approach if not confident
- Extend SAP IQ's ability to provide filtered results prior to optimization

Approach

We perform index-based adaptive sampling (IBAS) when uniformity or independence assumptions are required by constant-time estimation

- n:m joins
- 1:n with filtering on the primary key source (depicted below). Under certain conditions, SAP IQ can provide the query optimizer with the results of filter predicates

1. Obtain result set of predicate filters on the primary key source
2. Generate a sample on the result set. The sample size is fixed for a configurable number of rounds
3. Obtain the sum of counts of matching tuples in the foreign key source using persisted counts in B-Tree based indexes
4. Compute the current rounds estimate, update the inter-round variance metric
5. Perform a statistical quality test. If failed, determine the next sample size (the variance metric can be used) and got to 2



Evaluation

- Our benchmark is a subset of relevant queries from an industry-standard benchmark to evaluate IBAS **prototyped on SAP IQ**
- The datasets are generated with a moderate level of skew (Zipfian factor of 3) with varying sizes
- 3 individual runs of the benchmark under different configurations with a cold cache with execution times collected

IQ-1GB	HDL-1TB*	HDL-1TB-PKF
~1GB	~1TB	~1TB
On-prem IQ	HANA Cloud Data Lake	HANA Cloud Data Lake
Any estimation assumptions are targeted	Any estimation assumptions are targeted	Joins with filtering on only the primary key source
16% overall improvement	54% overall improvement*	57% overall improvement

Mean of the Improvements in Individual Queries

IQ-1GB	34%	-2.8%
HDL-1TB*	23%	-5.7%
HDL-1TB-PKF	31%	-3.6%

*optimizer hint required to correct bug under our estimates

IBAS can address the impact of filtering on join size estimation with relatively minimal overhead