Approximating User-Defined Functions in BlinkDB


BACKGROUND

SELECT avg(sessionTime) FROM Table WHERE city='Berkeley' GROUP BY dt, os, isp WITHIN 2 SECONDS

BlinkDB augments standard SQL-like queries to provide Response Time or Error/Confidence Interval Guarantees

Offline Sampling Module: BlinkDB maintains a set of multi-dimensional and multi-granular samples offline. These samples are either striped on hundreds of disks or cached in memory.

Per-Query Sample Selection Module: This module predicts per-query cost in BlinkDB and assigns an appropriately sized sample based on their error and/or response time requirements.

Error Bars and Confidence Intervals: All answers are then augmented with statistical measures of error bars and confidence intervals based on either statistical closed form measures (for AVG, SUM, COUNT, VARIANCE, PERCENTILES etc.) or Bag of Little Bootstraps technique (for arbitrary User-Defined Functions)

MAINTAINS UNIFORM & BIASED SAMPLES

MILP optimization picks the set of columns to stratify on within a storage budget

ERROR-LATENCY PROFILE

RELIABILITY OF THE BOOTSTRAP

Bootstrap is not always reliable. Verification on a per-query/per data-distribution basis by comparison with ground truth.

Facebook Queries (%)

RUNTIME DIAGNOSTICS

This verification is too slow. Good heuristic (Kleiner et al. 2012): Compare bootstrap with ground truth for small samples and ensure it improves with sample size.

DESIGNED FOR INTERACTIVITY

END-TO-END PERFORMANCE