Initialization-Free Online Aggregation with Adaptive Sampling

Zhiyi Huang[1],Weijie Zhao[1], Florin Rusu[1], Kai Zeng[2]
University of California Merced[1], Microsoft[2]
zhuang29, wzhao23, frusu@ucmerced.edu; kaizeng@microsoft.com

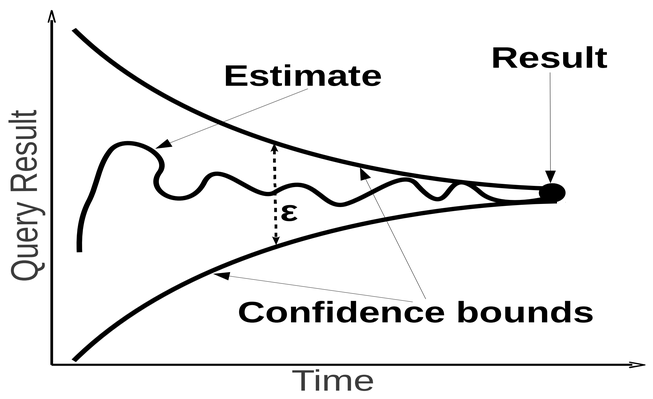
Motivation

- Analytics over big data is time-consuming and expensive
- Reliable estimates are necessary before a thorough investigation of the data
- Bootstrap provides statistical confidence bounds over samples
- Bootstrap over tuples requires either a preprocessing random shuffle or extracting a sample without replacement (both expensive)

Approach

- Bootstrap over block-level samples which has minimal overhead and no initialization
- Guide sampling and joins by collecting statistics at runtime
- Determine the randomness of the data and build adaptive estimators
- Return a bootstrap-based estimate if data is random or a block-level sample estimate if data is sorted

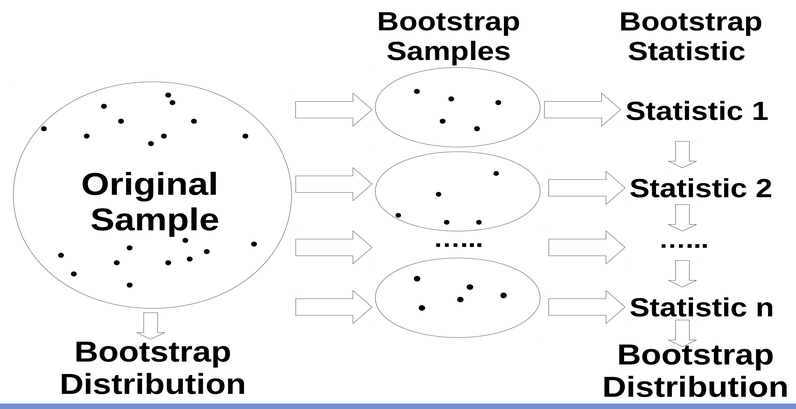
Online Aggregation



Block- vs Tuple-Level Sampling

- Tuple-level sampling produces tighter confidence bounds but requires shuffling or sampling at a deeper granularity
- Block-level sampling guarantees estimation accuracy independent of the data characteristics as long as blocks are processed in a random order

Bootstrap



A Bootstrap Example

R=

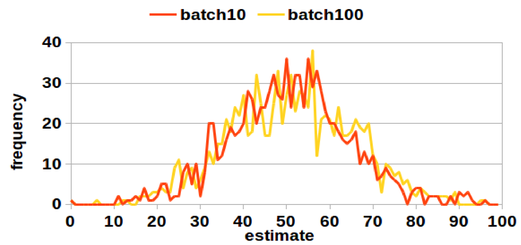
	l_pk	l_ex
t1	p1	10
t2	p2	50
t3	p2	20
t4	p3	40
t5	p4	30

$\hat{R}$=

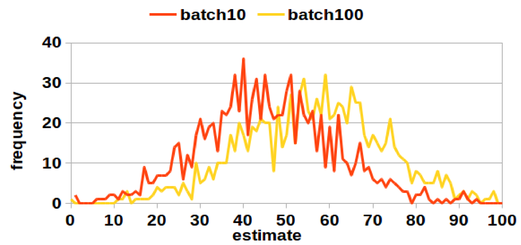
	l_pk	l_ex
t1	p1	10
t1	p1	10
t2	p2	50
t2	p2	50
t4	p4	40

	SUM(l_ex)
R	150
$\hat{R}$	150

Bootstrap Distribution

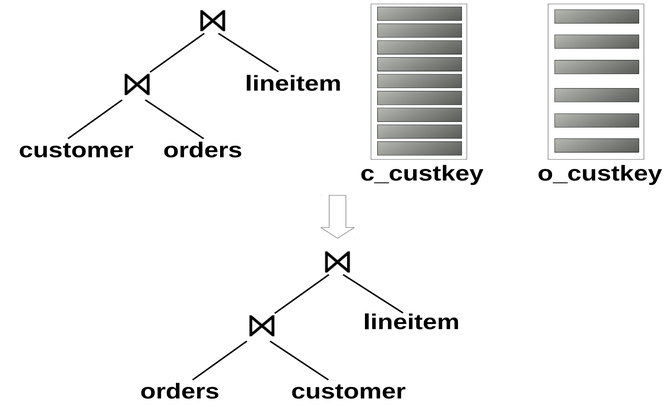


Bootstrap Distribution on Randomized Data



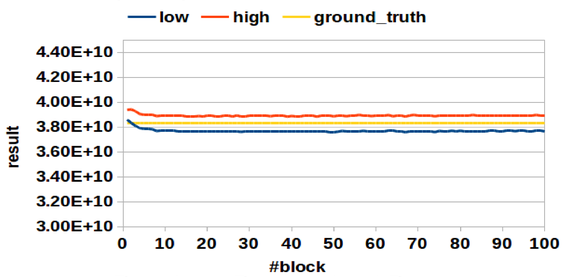
Bootstrap Distribution on Sorted Data

Multiple-Table Join

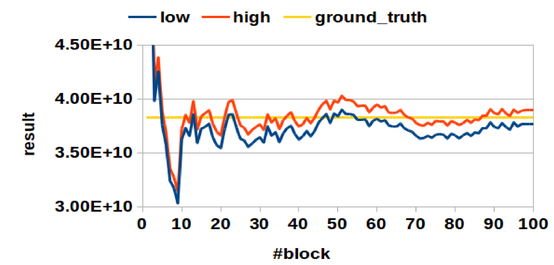


Experiments and Results

Query: `SELECT SUM(l_ex) FROM lineitem`



Result on Randomized Data



Result on Sorted Data