Approximation with Error Bounds in Spark

Guangyan Hu
Dept. of Computer Science

2/21/2019 at 01:00 pm
CoRE A (301)

Abstract

We introduce a sampling-based approximation framework for Spark, and show how multi-stage cluster sampling theories with population estimation techniques can be used to estimate error bounds for the approximate computations. We also show how adaptive stratified reservoir sampling can be used to avoid (or reduce) key losses in the final output. We evaluate a prototype implementation called ApproxSpark and our results show that (i) partition sampling can lead to greater reduction in execution time than data item sampling but lead to significantly larger error bounds (ii) stratified sampling reduces key loss and leads to more consistent error bounds across keys.

Examination Committee: Prof. Thu Nguyen (Chair), Prof. Ulrich Kremer, Prof. Yongfeng Zhang and Prof. Martin Farach-Colton