Experimental Evaluation of Sketching Techniques for Big Spatial Data

A. B. Siddique and Ahmed Eldawy
{msidd005,eldawy}@ucr.edu
Dept. of Computer Science and Engineering, University of California, Riverside

ABSTRACT

Ubiquitously connected devices, e.g., Internet of Things (IoT), space telescopes, social networks, and GPS-enabled gadgets, are contributing to the perpetual and swift growth of the data. 2.5 exabytes of daily-produced data, of which $60 - 80\%$ is geo-referenced. Space telescopes broadcast about 140 GB of data weekly. Availability of such large amount of data calls for new scalable query processing techniques. One of the techniques that is getting attention is *sketching* which summarizes the data and computes an approximate answer on the sketch. This general technique is used in partitioning [3], clustering [1], selectivity estimation [2], and visualization [4], among others. This paper introduces a sketching-based framework for big spatial data which provides four sketching methods and uses them to implement three common operations, namely, partitioning, clustering, and selectivity estimation. The framework is executed in three phases, sketching, local operation, and generalization, which can apply to a wide range of operations on big spatial data.

Sampling is a widely used sketching technique, but there exist other techniques such as uniform and non-uniform histograms which are not well-studied due to two challenges. First, each sketching method has a different representation and creation parameters, e.g., sampling ratio or number of histogram cells, which make it hard to compare their performance. Second, while existing algorithms can be used as-is with samples, other sketching methods might require some tweaks to the algorithms to work. This work provides a comprehensive evaluation to understand the trade-offs in the different sketching techniques for big spatial data.

In this paper, we present a three-phase sketching-based framework for big data processing. The first phase uses Spark to efficiently compute four types of data sketches, namely, sampling, uniform, non-uniform, and enhanced histograms. To make the sketching methods comparable, we define a parameter B which indicates the memory budget. Regardless of their representation, all sketching methods are designed to use up-to that memory budget. The second phase uses a single-machine to process the sketch and provide a partial answer to three popular and diverse operations, namely, partitioning, clustering, and selectivity estimation. Previous work mostly applied these techniques with sample-based sketches except for selectivity estimation which also used histograms. In this paper, we propose histogram-based spatial partitioning and K-means clustering and show that they can outperform sampling-based methods. The third phase takes the partial answer and scans all the data in parallel to generalize the answer to the entire dataset.

In our experiments, we use both real and synthetic datasets of up-to 2.7 billion records and 100 GB of data. We vary the memory budget that we use for sketching and study its effect in both the execution time and quality of the results.

CCS CONCEPTS

• **Information systems** → **Clustering**; **Summarization**; • **Theory of computation** → *Sketching and sampling*; Data structures and algorithms for data management;

KEYWORDS

Sketching, Clustering, Partitioning, Selectivity Estimation

REFERENCES

[1] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. 2012. Scalable k-means++. *Proceedings of the VLDB Endowment* 5, 7 (2012), 622–633.

[2] Harry Chasparis and Ahmed Eldawy. 2017. Experimental evaluation of selectivity estimation on big spatial data. In *Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data*. ACM, 8.

[3] Ahmed Eldawy, Louai Alarabi, and Mohamed F Mokbel. 2015. Spatial partitioning techniques in SpatialHadoop. *Proceedings of the VLDB Endowment* 8, 12 (2015), 1602–1605.

[4] Yongjoo Park, Michael J. Cafarella, and Barzan Mozafari. 2016. Visualization-aware sampling for very large databases. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016.* 755–766. DOI: http://dx.doi.org/10.1109/ICDE.2016.7498287

SoCC'18, October 11–13, 2018, Carlsbad, CA USA

ACM ISBN 978-1-4503-6011-1/18/10.
https://doi.org/10.1145/3267809.3275464