OmniSketch: Efficient Multi-Dimensional High-Velocity Stream Analytics with Arbitrary Predicates

W. R. Punter [1], O. Papapetrou [1], and M. Garofalakis [2]

[1] Eindhoven University of Technology, The Netherlands
{w.r.punter,o.papapetrou}@tue.nl
[2] Athena Research Center, Greece
minos@athenarc.gr

A key need in different disciplines is to perform analytics over fast-paced data streams, similar in nature to the traditional OLAP analytics in relational databases – i.e., with filters and aggregates. Storing unbounded streams, however, is not a realistic, or desired approach due to the high storage requirements, and the delays introduced when storing massive data. Accordingly, many synopses/sketches have been proposed that can summarize the stream in small memory (usually sufficiently small to be stored in RAM), such that count aggregates can be efficiently approximated, without storing the full stream. However, past synopses predominantly focus on summarizing single-attribute streams, and cannot handle filters and constraints on arbitrary subsets of multiple attributes efficiently. In this work, we propose, analyze, and evaluate a novel sketching tool, termed OmniSketch[1], that effectively addresses both space and time efficiency by combining sketching with sampling. OmniSketch, combines the compactness of sketches, which is necessary for reducing the memory constraints, with the generality of sampling, which is key for supporting general queries, on predicates that are dynamically decided at query time. In a nutshell, an OmniSketch for summarizing a p-attribute data stream consists of p individual small-memory sub-sketches, each similar to a Count-Min sketch. However, unlike Count-Min sketches, the cells in the OmniSketch sub-sketches contain fixed-size summaries of all records that hash into them. At query time, the sub-sketches that are relevant to the query, and the relevant cells from each sub-sketch, are located and queried to estimate the answers. Unlike previous work[2], OmniSketch offers computational complexity (for both updates and queries) that scales linearly with the number of attributes – instead of exponentially – rendering it the only viable, general-purpose solution, to date, for summarizing fast-paced streams with many attributes in small space. Our sketch is backed by a theoretical analysis for providing formal error guarantees, and an automated initialization algorithm that builds on the theoretical analysis to fully utilize the available sketching memory. We evaluate OmniSketch experimentally on both real and synthetically-generated streams, and compare it with Hydra[2], the state-of-the-art competitor. Our experiments confirm that OmniSketch is the only viable option for summarizing complex streams, and comes with a favorable complexity-accuracy tradeoff.

[1] Punter, W.R. and Papapetrou, O. and Garofalakis, M. (2023). OmniSketch: Efficient Multi-Dimensional High-Velocity Stream Analytics with Arbitrary Predicates. *arXiv preprint arXiv:2309.06051*

[2] Manousis, A. and Cheng, Z. and Basat, R.B. and Liu, Z. and Sekar, V. (2022) Enabling Efficient and General Subpopulation Analytics in Multidimensional Data Streams. *Proc. VLDB Endow.* 15, 11 (jul 2022), 3249-3262.