SWAG: An Experimental Study of Exact and Approximate Window-Based Non-Incremental Aggregations

Advisors

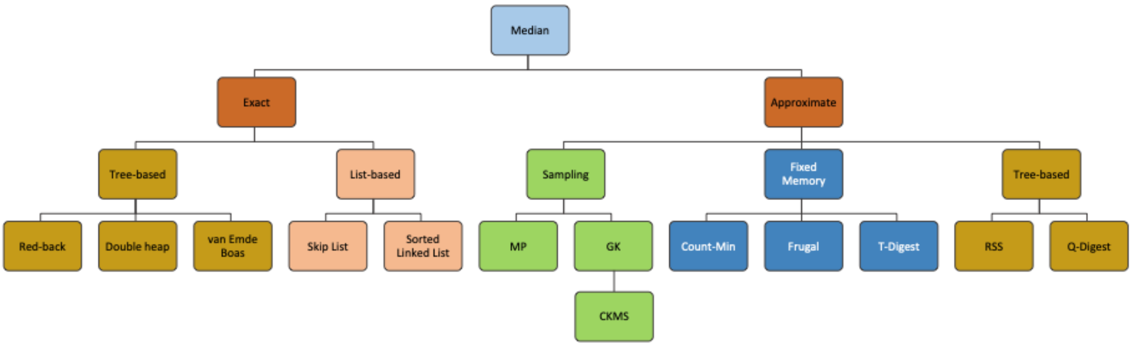
- Riccardo Tommasinii (INSA Lyon riccardo.tommasini@insa-lyon.fr)
- Angela Bonifati (Lyon 1 Univ., angela.bonifati@univ-lyon1.fr)

In addition, the project will be carried on in collaboration with Prof. Ahmed Awad from the University of Tartu, Estonia.

Context

Computing aggregations is one of the most common applications on data streams. Moving averages of trade markets, temperature sensor readings, and heart-beat monitoring are just a few examples of well-known scenarios where a user needs to receive moving averages, maximum values, and quantiles of the incoming data. In practice, sliding windows are the most common approach to slicing an unbounded data stream. Thus, several algorithms have been developed to compute aggregates over such windows efficiently. However, most of these algorithms focus only on aggregates that can be incrementally computed (e.g., sum, average, count). Little to nothing is known about the performance of non-incremental aggregates (e.g., median and quantiles) in general. In this context, key efficiency factors include incremental, pre-aggregation, and sharing of values among overlapping windows.
Moreover, in some instances with memory restrictions, sketch-based and approximate techniques are used to meet this constraint by sacrificing accuracy. In this paper, we present a comprehensive evaluation for "sum-like," "max-like," and "median-like" aggregations concerning performance and value calculation, i.e., exact vs. approximate.

Objectives



The goal of this project is to design an experimental study for non-incremental aggregations over streams. The project requires

- implementing state-of-the-art algorithms (see figure above) on top of Apache Flink, the state-of-the-art stream processing system, and Scotty, a recent library for efficient window aggregate computation.
- designing experiments for a systematic evaluation using both synthetic and real-world data.

Methodology

- Familiarise yourself with the concepts related to stream processing and streaming analytics
- Familiarise yourself with the distributed stream processing engine Flink and the library Scotty
- Familiarise yourself with a cloud environment for executing experiments in distributed settings
- Design and execute extensive experimentation measuring throughput, latency, and memory consumption for the selected aggregation algorithms.

References

[1] C. C. Aggarwal and P. S. Yu. A survey of synopsis construction in data streams. In C. C. Aggarwal, editor, Data Streams - Models and Algorithms, volume 31 of Advances in Database Systems, pages 169–207. Springer, 2007.

[2] J. Astola and T. G. Campbell. On the computation of the running median. IEEE Trans. Acoustics, Speech, and Signal Processing, 37(4):572–574, 1989.

[3] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas. Apache flinkTM: Stream

[4] S. Chintapalli, D. Dagit, B. Evans, R. Farivar, T. Graves, M. Holderbaugh, Z. Liu, K. Nusbaum,K. Patil, B. Peng, and P. Poulosky. Benchmarking streaming computation engines: Storm, flink and spark streaming. In 2016 IEEE International Parallel and Distributed Processing Symposium Workshops, IPDPS Workshops 2016, Chicago, IL, USA, May 23-27, 2016, pages 1789–1792. IEEE Computer Society, 2016.

[5] M. Hirzel, S. Schneider, and K. Tangwongsan. Sliding-window aggregation algorithms: Tutorial. In Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems, DEBS 2017, Barcelona, Spain, June 19-23, 2017, pages 11–14. ACM, 2017.