Stream Processing on Modern Hardware

(Master project, Winter 2021/22)

Current Stream Processing Engines (SPEs) can process millions of records across hundreds of nodes to analyze an ever growing amount of real-time data. However, the most widely used SPEs, such as Apache Flink, Spark Streaming, or Storm are all JVM-based and do not utilize the servers' hardware efficiently. Recent work [1] shows that many hardware-specific optimizations can be made to improve the efficiency of each individual node (i.e., scale-up instead of scale-out). These optimizations include, e.g., query compilation, use of modern storage devices, or use of RDMA.

We are currently developing a new SPE for modern hardware at our group, which tackles some of these challenges. In this project, you will actively develop components of this new system, which we plan to open source soon. The project is still in very early stages, so there are many topics to choose from, depending on your interests. Some ideas are: efficient data distribution between operators via RDMA, efficient hardware-optimized streaming joins and aggregations, efficient checkpointing and recovery, efficient query compilation via architecture-aware optimizations, query plan optimization, and many more.

Students will learn the inner workings of stream processors and data management systems in general, with a particular focus on query compilation and modern hardware. It is targeting students interested in acquiring skills in data management, stream processing, data flows, compilers, and low-level systems programming. The project will be implemented in C++.

Information on query compiled SPEs can be found in the recent SIGMOD '20 paper Grizzly [2]. General information and an introduction on stream processing can be found in the O'Reilly blog posts by Tyler Akidau [3, 4] and the stream processing book [5].

Grading

Courses applicable: ITSE (Masterprojekt), DE (Data Engineering Lab)

Graded activity:

- Implementation / group work
- Final report (8 pages, double-column, ACM-art 9pt conference format)
- Final presentation (20 min)

Contact

Lawrence Benson

Literature

[1]: Zeuch et al., 2019. Analyzing Efficient Stream Processing on Modern Hardware, *PVLDB*
[2]: Grulich et al., 2020. Grizzly: Efficient Stream Processing Through Adaptive Query Compilation, *SIGMOD*
[3]: Tyler Akidau: Streaming 101. https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101
[4]: Tyler Akidau: Streaming 102. https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-102
[5]: Tyler Akidau, Slava Chernyak, Reuven Lax: Streaming Systems. O'Reilly. http://streamingsystems.net/