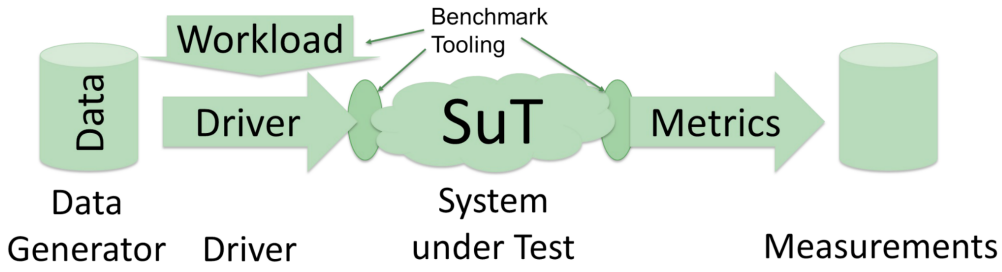
Benchmarking Real-Time Analytics Systems

(Master project, Winter 2023)

With ever growing amounts of data, analyzing it in real-time is becoming increasingly relevant. Unlike classical data warehouses that ingest data in batch, novel real-time analytical engines aim to process data as soon as it arrives to allow for analysis on fresh data. Compared to traditional OLAP systems, real-time systems are still rapidly developing. To understand the performance of OLAP systems, many benchmarks exist (e.g., TPC-H [1], TPC-DS [2], SSB [3]), but none of them consider data freshness and streaming ingest as supported by these novel systems. On the other end of the spectrum are streaming benchmarks, which evaluate stream processing engines, but do not consider analytical queries (e.g., LinearRoad Benchmark [4], Yahoo Streaming Benchmark [5]). To better understand the limitations and advantages of current systems, we plan to develop a real-time analytics benchmark and compare different real-time analytics systems.

An important step in developing such a benchmark is to see how it is implementable in various system types and what the performance characteristics are. To gain insights into this, the master project should implement an early version of the benchmark in a few *systems under test*, e.g., Postgres, DuckDB, Apache Flink, Apache Druid. Each of these systems has a different focus and supports real-time analytics in varying degrees.



As a basis for the benchmark specification and tooling, we will reuse existing benchmarks, particularly the Yahoo Streaming Benchmark [5], TPCx-IoT [6], and SSB [3], which cover subsets of our requirements.

Participants will learn about current trends in real-time analytics and how to apply them to large-scale open-source projects. They will gain deeper insights into the selected systems by implementing a high-performance benchmark in them. Students should be interested in data management, performance management, and comfortable in programming in SQL, C++, and/or Java, depending on the system and interface.

Grading

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

Graded activity:

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

Contact

Tilmann Rabl, Lawrence Benson

References

1. TPC-H - https://www.tpc.org/tpch/default5.asp
2. TPC-DS - https://www.tpc.org/tpcds/default5.asp
3. Star Schema Benchmark - https://www.cs.umb.edu/~poneil/StarSchemaB.PDF
4. Linear Road Benchmark - https://www.cs.brandeis.edu/~linearroad/
5. Yahoo Streaming Benchmark - https://yahooeng.tumblr.com/post/135321837876/benchmarking-streaming-computation-engines-at
6. TPCx-IoT - https://www.tpc.org/tpcx-iot/default5.asp