CS3550-ADMT25 Guests Lectures

Date: Mar 9, 2021

Title: A Holistic View of Stream Partitioning Costs & Expediting Stream Processing with Accuracy Guarantees

Abstract: Stream processing has become the dominant processing model for monitoring and real-time analytics. Modern Parallel Stream Processing Engines (pSPEs) have made it feasible to increase the performance in both monitoring and analytical queries by parallelizing a query's execution and distributing the load on multiple workers. A determining factor for the performance of a pSPE is the partitioning algorithm used to disseminate tuples to workers. Until now, partitioning methods in pSPEs have been similar to the ones used in parallel databases and only recently load-aware algorithms have been employed to improve the effectiveness of parallel execution. We identify and demonstrate the need to incorporate aggregation costs in the partitioning model when executing stateful operations in parallel, in order to minimize the overall latency and/or throughput. Towards this, we propose new stream partitioning algorithms, that consider both tuple imbalance and aggregation cost. We evaluate our proposed algorithms and show that they can achieve up to an order of magnitude better performance, compared to the current state of the art.

Further, Stream Processing Engines (SPEs) tend to overprovision resources, and online scaling is required in order to overcome overloaded situations. Current attempts for expediting stateful processing are impractical, due to their inability to uphold the quality of results, maintain performance, and reduce memory requirements. In this talk, we present the SPEAr system, which can expedite processing of stateful operations automatically by trading accuracy for performance. SPEAr detects when it can accelerate processing by employing online sampling and accuracy estimation at no additional cost. We built SPEAr on top of Storm and our experiments indicate that it can reduce processing times by more than an order of magnitude, use more than an order of magnitude less memory, and offer accuracy guarantees in real-world benchmarks.

Speaker: Dr. Nikos R. Katsipoulakis is an alumni of Pitt CS and the ADMT Lab. Nikos is a member of the AWS Redshift Query Processing team. His work focuses on Metadata management.