Caching Algorithms for ViDa

Keywords: In-situ query processing, Caching, Recycling Materialized Results

Problem: Scientists in many disciplines generate massive amounts of data. The amounts of data produces are already so big that they can barely be managed. In-situ query processing [1] offers flexibility and interactivity for scientists to query directly over raw data files and offers performance comparable to state-of-the-art relational DBMS. ViDa [2] is the latest version of our in-situ query processing system. One of the key ideas driving the performance of this system is caching. By caching and then reusing previously computed query operator results, the system is able to achieve significant performance gains. However, the current caching infrastructure is somewhat ad-hoc and not well-understood. Using more intelligent caching algorithms that adapt to the differences in costs and benefits of various data formats and sizes can help further improve the querying performance of ViDa over diverse data sources.

Project: The goal of this project is to design, implement and evaluate effective cost-aware caching algorithms for Vida. By taking costs and benefits into account, the algorithms should enable improved querying performance over a variety of raw input formats.


Plan:
1. Get acquainted with the literature.
2. Understand RAW/ViDa code.
3. Design and implement caching modules, data structures and algorithms.
4. Run experiments to evaluate performance gains.

Supervisor: Prof. Anastasia Ailamaki, anastasia.ailamaki@epfl.ch
Responsible collaborator(s): Tahir Azim, tahir.azim@epfl.ch
Duration: 4 months