Design and implement primitives for parallel in-situ query processing

**Keywords:** stageDB, Thread Building Blocks

**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 straight over raw data files and offers performance comparable to state-of-the-art relational DBMS. Most existing approaches for in-situ query processing do not offer parallel execution and those who do offer [2], their programming model does not allow many optimizations or changes in the execution plan.

A current trend towards parallelization are Thread building blocks (TBB). They allow programs to take advantage of multi-core processors. Through different techniques (i.e., work stealing) TBB balance processing among multiple cores.

**Project:** The goal of this project is to create a system which integrates the two approaches described above. Create an infrastructure of primitive objects (e.g., Thread building blocks) which will allow to build an in-situ query engine that will scale up gracefully. The main idea is to simplify the correlation between the necessary objects and enable parallelization.

[1] **NoDB: Efficient Query Execution on Raw Data Files.** Ioannis Alagiannis, Renata Borovica, Miguel Branco, Stratos Idreos and Anastasia Ailamaki.

[2] **SCANRAW: A Database Meta-Operator for Parallel In-Situ Processing and Loading.** Yu Cheng, Florin Rusu

**Plan:**
1. Get acquainted with the literature and TBBs
2. Design necessary primitives (descriptions and API)
3. Design system that will combine those primitives (pipelines, descriptions language)
4. Implementation

**Supervisor:** Prof. Anastasia Ailamaki, anastasia.ailamaki@epfl.ch

**Responsible collaborator(s):** Manos Karpathiotakis, manos.karpathiotakis@epfl.ch
Matthaios Olma, matthaios.olma@epfl.ch

**Duration:** 3 months