DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

EPFL

Machine Learning Techniques for Performance Prediction in Modern Analytical Engines

Keywords: Query Optimization, Machine Learning, Query Plans

Problem: Analytical DBMS engines incorporate a crucial component called as *query optimizer* to pick an efficient query plan for execution. In order to find the optimal plan, we need an accurate cost model. An accurate cost model also helps in predicting the performance (execution time) of a query plan. Predicting query execution time is increasingly important in the context of databases as a service, along with system management decisions such as admission control, query scheduling and progress monitoring.

Project: The goal of this project is to use machine learning techniques for Proteus [2], an in-house DBMS engine for analytical queries. Proteus currently uses Calcite [1] which is a cost-based optimizer to generate plans. The student needs to evaluate the performance of state-of-the-art machine learning techniques which exist for relational database systems by adopting it to Proteus (one such attempt for RDBMS can be seen in [3]). To start with, the student can assume a simplistic model of the engine and evaluate its performance. Then, expand the scope of the earlier model to handle the practical scenarios. The outcome of the project is to propose the *best* machine learning technique (either from literature or a new one) for the above mentioned problems which outperforms the baseline (i.e., Calcite).

Plan:

1. A literature survey of machine learning techniques for RDBMS
2. Implement and evaluate the state-of-the art RDMS technique for Proteus using simplistic model
3. Extend Step 2 to handle practical scenarios

Supervisor: Prof. Anastasia Ailamaki, anastasia.ailamaki@epfl.ch
Responsible collaborator(s): Srinivas Karthik, srinivas.venkatesh@epfl.ch
Duration: 10 Weeks

References

[1] E. Begoli et al. Apache calcite: a foundational framework for optimized query processing over heterogeneous data sources. In SIGMOD, pages 221–230, 2018

[2] M. Karpathiotakis, I. Alagiannis, and A. Ailamaki. Fast Queries Over Heterogeneous Data Through Engine Customization. In PVLDB, pages 972-983, 2016

[3] Marcus et al. Plan-Structured Deep Neural Network Models for Query Performance Prediction. In PVLDB, pages 1733-1746, 2019