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/

Data-less Distributed Query Processing using ML Models

Keywords: Approximate Query Processing, Spark, Machine Learning

Problem: Due to the exponential growth of the data volume, it is prohibitively expensive to compute the exact result of analytical queries. For this reason, query processing techniques that compute approximate answers in sublinear time have been proposed. Recently, machine learning models have been used as a compact but accurate alternative to sampling-based approaches. However, research in this direction is still in its first steps and has not provided definitive answers on how to train the models, considering the distributed nature of the data and the high cost of data movement.

Project: The goal of the project is to design a framework for distributed training and inference of the ML models used for approximate query processing. The student is expected to apply distributed machine learning concepts to train the models while keeping the data mostly in-place and without extensive shuffling between the servers. To build the prototype, distributed frameworks such as Spark will be used.

Plan:

1. Study relevant literature in approximate query processing using ML models and reproduce the results in a single-node.
2. Develop a distributed method for the training and inference of the ML models.
3. Evaluate the performance gains and the convergence penalties compared to the single-node baseline.

Supervisor:	Prof. Anastasia Ailamaki, anastasia.ailamaki@epfl.ch
Responsible collaborator(s):	Sioulas Panagiotis and Sanca Viktor
Duration:	1 semester

References

1. S. Agarwal, BlinkDB: queries with bounded errors and bounded response times on very large data. In EuroSys 2013.
2. Q. Ma and P. Triantafillou, DBEst: Revisiting Approximate Query Processing Engines with Machine Learning Models. In SIGMOD 2019.