Data-less Distributed Query Processing using Sketches

**Keywords:** Approximate Query Processing, Spark, Sketches, Count-Min Sketch

**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. While machine learning models have been proposed as an alternative to sampling-based approaches, sublinear methods such as sketches (e.g. count-min sketch) have long been used to estimate aggregations over large volumes of data. Being cheap to construct and maintain, sketches are a promising and theoretically sound replacement for proposed ML 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 efficient distributed data sketch creation and inference for approximate query processing. The student is expected to understand and use data-sketch properties to reduce extensive data shuffling between the servers, and understand and evaluate the tradeoffs between accuracy and reduced space/time complexity. To build the prototype, distributed frameworks such as Spark will be used.

**Plan:**
1. Study relevant literature in approximate query processing using sketches and reproduce the results in a single-node.
2. Develop a distributed method for maintaining and using sketches.
3. Evaluate the performance gains compared to the accurate, and sampling method baseline.
4. 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**

2. G. Cormode and S. Muthukrishnan, An improved data stream summary: the count-min sketch and its applications