University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, 25. Januar 2023

MSc Thesis
Topic: End-to-End Implementation of Pair-Wise Correlation Computation in a Streaming System

Due to the ubiquity of sensors and sensor networks, the monitoring of time series data has become more and more important. Given a number of data streams, one particular task is to determine all pairs of data streams that are currently correlated (i.e., the most recent windows of the data streams are correlated). Dimensionality-reduction filter-and-refine techniques have been developed to speed up the process of finding these pairs efficiently. Since there is a considerable number of different techniques that have never been directly compared to each other, it would be interesting to benchmark the algorithms against each other in a common and realistic environment.

The aim of this thesis is to implement different dimensionality-reduction techniques in a streaming platform so that they can be benchmarked under realistic conditions. The work is structured roughly into the following tasks:

Tasks

T1: Understanding the algorithms
The first task is to study the algorithms to gain a deeper understanding of how they work. This includes looking at an existing implementation of the algorithms in R. In particular, these algorithms are Discrete Fourier Transform (DFT), Singular Value Decomposition (SVD), Piecewise Aggregate Approximation (PAA), Piecewise Linear Approximation (PLA), and Chebyshev Polynomials (CHBV). As a baseline, incremental Pearson needs to be implemented.

T2: Pick a streaming system
In order to be able to run the benchmarks under realistic conditions, the algorithms

should all be implemented in an industrial-strength streaming system, such as Flink, Kafka, or Spark. The goal of this task is to identify a streaming platform that is well-suited for the benchmarking by conducting a brief feasibility study, e.g., by implementing a few prototypes in each system.

T3: Implementing the algorithms
This is the main task of the thesis. The aim is to implement the algorithms listed under task T1 in the system chosen in task T2. The programming language will depend on the chosen system.

T4: Setting up and running benchmarks (optional)
Some small benchmarks will already be run in task T3 to test the correctness of the implementation. The goal of this task is to develop a benchmark framework that runs the implemented algorithms on several real-world datasets and outputs the results of the benchmark runs.

T4: Summarizing the findings in a report/thesis
At the end of the project, a report/thesis needs to be written up.

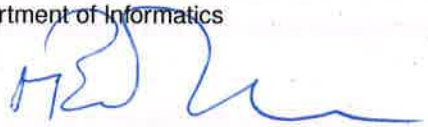
T5: Presentation
Present the thesis in a DBTG meeting (25 minutes presentation).

Supervisor: Sven Helmer (helmer@ifi.uzh.ch)

Start date: 1 February 2023

End date: 31 July 2023

University of Zurich
Department of Informatics



Prof. Dr. Michael Böhlen
Professor