MineLiB: A Library of Sub-Operators for Data Mining Algorithms

Accelerating data mining algorithms on modern computing hardware is critical for efficient use of the data deluge available in the cloud today.

The abstraction level used in writing data mining algorithms is either high (e.g., using a DSL and reyling on the compiler to generate platform efficient code), or low (e.g., using platform specific programming languages like C, VHDL, CUDA, etc.). These abstraction levels make it hard for an optimizer to schedule parts of the algorithm on the different processing units of hybrid hardware platforms. Moreover, many of the optimization decisions are then hardly transferable to other algorithms or across different platforms.

Specifying an intermediate abstraction level of algorithmic building blocks (*sub-operators*) which can be shared among a wide range of operations, significantly simplifies the job of an optimizer and can make the portability of the algorithms a lot easier. Each sub-operator can then be executed on a variety of processing units (CPU, GPU, FPGA, etc.). The idea of using sub-operators has been recently explored for relational database operators [1,2], to facilitate better optimization of database queries and cross-platform execution.

This project adopts the abstraction level of sub-operators to express and optimize data mining algorithms on modern hardware. The goal is the development of a library of sub-operators tailored for data mining algorithms (MineLiB as shown in Fig. 1). The deliverables are a set data types, kernel building blocks (code fragments), their models, and the corresponding API which can be used to fully express the algorithm's content.

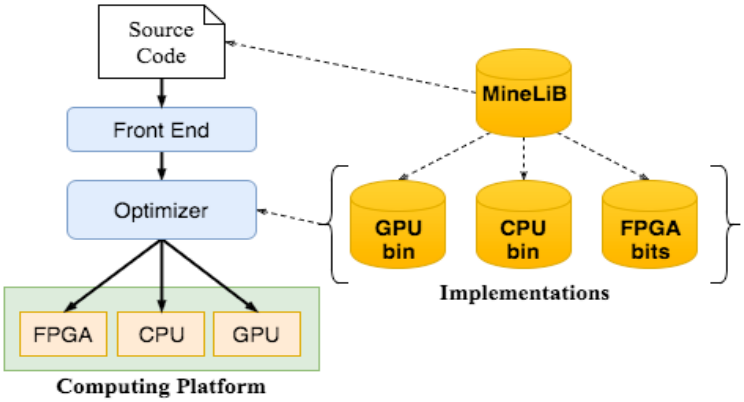

Figure 1: A depiction of a compilation flow using MineLiB.

The main challenge is finding the appropriate abstraction level, and selecting a proper set of sub-operators. On the one side, the set should be expressive enough to be able to construct a wide range of algorithms. On the other side, the internal design and implementation of each sub-operator should also account for efficient integration on variety of hardware platforms.

The proposed project work flow includes the following tasks:

- Study a range of data mining algorithms, such as the ones in MineBench [3].
- Define the properties and features of the set of sub-operators, such that they benefits from various characteristics of modern hardware.
- Propose a set of sub-operators and data types and define a programming interface for it.
- Through a few study cases, demonstrate the benefit of using the proposed sub-operators library by implementing and optimizing a few sub-operators on CPU or FPGA (or both).

If you are interested in this project, please contact Muhsen Owaida (mohsen.ewaida@inf.ethz.ch) or Jana Giceva (gicevaj@inf.ethz.ch). This project will be under the supervision of Prof. Gustavo Alonso.

[1] Holger et. al, "Voodoo - A Vector Algebra for Portable Database Performance on Modern Hardware", VLDB 2016.
[2] He et a., "Relational query coprocessing on graphics processors", TODS 2009.
[3] Narayanan et. al, "MineBench: A Benchmark Suite for Data Mining Workloads", IISWC 2006.