Holistic Extensibility for Integrated Data Analysis Pipelines in DAPHNE (Invited Talk)

Patrick Damme

Postdoctoral Researcher, Technische Universität Berlin

Abstract

Integrated data analysis (IDA) pipelines combining data management and query processing, machine learning, and high-performance computing become increasingly important in practice. Deploying IDA pipelines is cumbersome and quickly leads to overheads and hardware under-utilization due to the suboptimal integration of existing systems. At the same time, today's hardware challenges lead to increasing specialization in terms of compute and storage devices as well as the data representation, which further complicates the deployment. To address these issues, in the DAPHNE project, we build an open and extensible system infrastructure for IDA pipelines, aiming for increased productivity and performance. End users define IDA pipelines in a domain-specific language offering a unified experience over relational and linear algebra, augmented by high-level data science primitives. Internally, an IDA pipeline runs through an optimizing compiler and is efficiently executed in local or distributed environments exploiting heterogeneous hardware and computational storage. Researchers may easily extend many aspects of DAPHNE including custom data representations, operator kernels, optimizer passes, and runtime schedulers to experiment with new specialized approaches in a full-fledged system. In this talk, I will give an overview of our ongoing work in DAPHNE and on its holistic extensibility in particular.

Speaker Biography: Patrick Damme is a postdoc researcher in the DAMS Lab research group at TU Berlin and the BIFOLD in Berlin, Germany. His research interests are centered around database systems, machine learning systems, and techniques for making complex analyses of large data volumes efficient, scalable, and simple. He is one of the main contributors to DAPHNE, an open and extensible system infrastructure for integrated data analysis pipelines, which he started working on in his previous occupation as a postdoc at Graz University of Technology and the co-located Know-Center in Graz, Austria. Patrick received his PhD in computer science from TU Dresden, Germany, where he was a part of the Dresden Database Systems Group.

Joint Workshops at 49th International Conference on Very Large Data Bases (VLDBW'23) — Second International Workshop on Composable Data Management Systems (CDMS'23), August 28 - September 1, 2023, Vancouver, Canada