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, August 28, 2022

Bachelor Thesis (18 ECTS)
Database Technology

Topic: Integration of Matrix Transposition into Database Systems

Linear algebra expressions have become very popular and are frequently being used for data analysis and machine learning. Examples of widely used linear algebra expressions are the matrix cross product X^TX (i.e., the Gram matrix) and the least square estimator $(X^TX)^{-1}X^Ty$. The implementation of such expressions over a database system backend is non-trivial. One of the main reasons is that matrix transposition, outer product, and singular value decomposition yield result relations with a schema that depends on the values in the argument relations. For instance the transpose operation yields a relation where the number of attributes is equal to the number of tuples of the argument relation.

Operations with a data dependent result schema are difficult to handle for database systems. A first reason is that the schema of the result is not known at compile time. This affects the entire query optimization and query evaluation pipeline since neither number nor names nor types of the attributes are known. A second reason is the scalability. Database system scale extremely well with the number of rows but the scalability in terms of numbers of columns is limited. Since a transposition exchanges rows and columns it usually yields a relation with a large number of columns that most state-of-the-art systems cannot handle efficiently.

The goal of this Bachelor thesis is to design, implement and empirically evaluate a solution to efficiently handle (sub)expressions that yield relations with data-dependent schemas. As a starting point use the Relational Matrix Algebra (RMA) [3], which reconciles relations and

matrices, and makes it possible to apply matrix operations to parts of a relation without losing contextual information. Integrate you solution into MonetDB.

The work includes the following tasks:

T1: Study related work about data management and machine learning systems that combine relational and linear algebra functionality [4, 3, 1, 2]. Analyze and describe cases that use transpose operations and identify expressions that are often used in data science pipelines.

T2: Propose a solution to store relations in their transformed form. Design and implement evaluation algorithms to efficiently compute expressions over relations that are possibly stored in their transformed form (for instance an algorithm to compute $\sigma(r)$ if r is stored in its transposed form). Integrate your solution into MonetDB. Pay attention to the handling of schema information and the scalability of your solution.

T3: Extend the query optimizer of MonetDB with optimization techniques that leverage the implemented evaluation algorithms. As an example, the MonetDB optimizer can be extended with an equivalence rule that rewrites an expression to an equivalent one without transposition.

T4: Empirically evaluate your solution on real world applications and realistic workloads.

T5: Write a thesis (approximately 50 pages).

References

[1] Matthias Boehm, Iulian Antonov, Sebastian Baunsgaard, Mark Dokter, Robert Ginthör, Kevin Innerebner, Florijan Klezin, Stefanie N. Lindstaedt, Arnab Phani, Benjamin Rath, Berthold Reinwald, Shafaq Siddiqui, and Sebastian Benjamin Wrede. Systemds: A declarative machine learning system for the end-to-end data science lifecycle. In *10th Conference on Innovative Data Systems Research, CIDR 2020, Amsterdam, The Netherlands, January 12-15, 2020, Online Proceedings*. www.cidrdb.org, 2020.

[2] Sudipto Das, Yannis Sismanis, Kevin Beyer, Rainer Gemulla, Peter Haas, and John Mcpherson. Ricardo: Integrating r and hadoop. pages 987–998, 06 2010.

[3] Oksana Dolmatova, Nikolaus Augsten, and Michael H. Böhlen. A relational matrix algebra and its implementation in a column store. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD 2020, online conference [Portland, OR, USA], June 14-19, 2020*, pages 2573–2587. ACM, 2020.

[4] Shangyu Luo, Zekai J. Gao, Michael N. Gubanov, Luis Leopoldo Perez, and Christopher M. Jermaine. Scalable linear algebra on a relational database system. *SIGMOD Rec.*, 47(1):24–31, 2018.