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

Master Thesis (30 ECTS)
Database Technology

Topic: Efficient in-place iterations in MonetDB

The enormous growth of stored data continues to challenge our ability to efficiently process and analyze data. Many analytical computations, e.g., Markov chain algorithms or various types of regressions, require a combination of operations from the relational algebra, operations form the linear algebra, and iterations. A number of attempts have been made to tackle selected combinations and aspects of these elements:

The Oracle UTL_NLA package [11] exposes selected BLAS and LAPACK operations on matrices represented as VARRAYs. SimSQL [10] is parallel relational database system that adds matrices as a new data type to support linear algebra operations. RMA [7] views relations as matrices enhanced with contextual information, and matrix operations can be applied to dynamically defined parts of relations. SciDB [4] is an array database with linear algebra as its main application. Rasdaman [1] is an array database system with declarative support for managing large n-D arrays. RIOT [5] makes R programs I/O-efficient with an array storage manager and an engine for statistical operations. Ricardo [6] supports R programming on top of Hadoop. SystemML [3] compiles R-like scripts into hybrid plans with local and distributed Spark operations. SystemDS [2] is an ML system for end-to-end data science with data integration, cleaning, preparation and model training, debugging and serving. AIDA [8] offers a system level integration of MonetDB and Python with a shared address space to limit data copying.

The goal of this Master thesis is to implement explicit iterations over relations for which modern database management systems do not offer efficient and scalable support. Current systems support explicit iterations through either recursive SQL queries or loops in UDFs [9, 12]. Such iterations, however, have not been designed for analytical purposes and the integration into query trees and the query optimizer is limited. Specifically, iterations are implemented as out-of-place loops that create new (temporary) tables in each step and have to repeatedly allocate and free memory. Clearly, this is a major limitations for the implementation of iterative algorithms.

The work includes the following tasks:

1. Extend MonetDB with iterations over relations:
   (a) Design and implement in-place iterations over relations. Focus on a solution that efficiently supports iterations that combine relational and linear algebra operations. The iterations shall be integrated into the query tree and the query optimization process, and its components shall be accessible to the optimizer.
   (b) Empirically evaluate the properties of the proposed solution.
2. Propose and implement representative cross-algebra optimization rules that include relational and linear algebra operations and iterations over relations.
3. Write a thesis (approximately 50 pages).
4. Present your thesis in a DBTG meeting.

References

[1] Peter Baumann, Dimitar Misev, Vlad Merticariu, and Bang Pham Huu. Array databases: concepts, standards, implementations. *Journal of Big Data*, 8(1):28, 2021.

[2] 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.

[3] Matthias Boehm, Michael Dusenberry, Deron Eriksson, Alexandre V. Evfimievski, Faraz Makari Manshadi, Niketan Pansare, Berthold Reinwald, Frederick Reiss, Prithviraj Sen, Arvind Surve, and Shirish Tatikonda. Systemml: Declarative machine learning on spark. *Proc. VLDB Endow.*, 9(13):1425–1436, 2016.

[4] Paul Brown. Overview of scidb: large scale array storage, processing and analysis. pages 963–968, 06 2010.

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

[6] Sudipto Das, Yannis Sismanis, Kevin Beyer, Rainer Gemulla, Peter Haas, and John

Mcpherson. Ricardo: Integrating r and hadoop. pages 987–998, 06 2010.

[7] 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.

[8] Joseph D'silva, Florestan Moor, and Bettina Kemme. Aida: abstraction for advanced in-database analytics. *Proceedings of the VLDB Endowment*, 11:1400–1413, 07 2018.

[9] Dimitrije Jankov, Shangyu Luo, Binhang Yuan, Zhuhua Cai, Jia Zou, Chris Jermaine, and Zekai Gao. Declarative recursive computation on an rdbms: or, why you should use a database for distributed machine learning. *ACM SIGMOD Record*, 49:43–50, 09 2020.

[10] 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.

[11] Oracle® Database. *PL/SQL Packages and Types Reference: UTL_NLA*. `https://docs.oracle.com/en/database/oracle/oracle-database/21/arpls/UTL_NLA.html`.

[12] Linnea Passing, Manuel Then, Nina C. Hubig, Harald Lang, Michael Schreier, Stephan Günnemann, Alfons Kemper, and Thomas Neumann. SQL- and operator-centric data analytics in relational main-memory databases. In *Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017*, pages 84–95. OpenProceedings.org, 2017.

University of Zürich
Department of Informatics

Prof. Dr. Michael Böhlen