Integrating analytics with relational databases

Raasveldt, M.

Bibliography

[1] Daniel Abadi, Samuel Madden, and Miguel Ferreira. Integrating Compression and Execution in Column-oriented Database Systems. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, pages 671–682, New York, NY, USA, 2006. ACM.

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[3] Rakesh Agrawal and Kyuseok Shim. Developing tightly-coupled data mining applications on a relational database system. In *In Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining*, pages 287–290. AAAI Press, 1996.

[4] Grant Allen and Mike Owens. *The Definitive Guide to SQLite*. Apress, Berkely, CA, USA, 2nd edition, 2010.

[5] Michael Armbrust, Reynold S. Xin, Cheng Lian, et al. Spark SQL: Relational Data Processing in Spark. In *Proceedings of the 2015 ACM SIGMOD International*

Conference on Management of Data, SIGMOD '15, pages 1383–1394, New York, NY, USA, 2015. ACM.

[6] M. M. Astrahan, M. W. Blasgen, D. D. Chamberlin, K. P. Eswaran, J. N. Gray, P. P. Griffiths, W. F. King, R. A. Lorie, P. R. McJones, J. W. Mehl, G. R. Putzolu, I. L. Traiger, B. W. Wade, and V. Watson. System r: Relational approach to database management. *ACM Trans. Database Syst.*, 1(2):97–137, June 1976.

[7] Jeff Bezanson, Stefan Karpinski, Viral B. Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. *CoRR*, abs/1209.5145, 2012.

[8] Peter Boncz, Marcin Zukowski, and Niels Nes. Monetdb/x100: Hyper-pipelining query execution. In *In CIDR*, 2005.

[9] Peter A. Boncz, Marcin Zukowski, and Niels Nes. Monetdb/x100: Hyper-pipelining query execution. In *CIDR 2005, Second Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2005*, pages 225–237, 2005.

[10] United States Census Bureau. American Community Survey. Technical report, 2014.

[11] Pierre Carbonnelle. Top IDE index. 2018.

[12] D. D. Chamberlin and R. F. Boyce. SEQUEL: A structured english query language. In *ACM SIGMOD*, page 249264, 5 1974.

[13] Surajit Chaudhuri and Kyuseok Shim. Optimization of queries with user-defined predicates. *ACM Trans. Database Syst.*, 24(2):177–228, June 1999.

[14] Qiming Chen, Meichun Hsu, and Rui Liu. Extend UDF Technology for Integrated Analytics. In TorbenBach Pedersen, MukeshK. Mohania, and AMin Tjoa, editors, *Data Warehousing and Knowledge Discovery*, volume 5691 of *Lecture Notes in Computer Science*, pages 256–270. Springer Berlin Heidelberg, 2009.

[15] Qiming Chen, Meichun Hsu, Rui Liu, and Weihong Wang. Scaling-up and speeding-up video analytics inside database engine. In SouravS. Bhowmick, Josef Kng, and Roland Wagner, editors, *Database and Expert Systems Applications*, volume 5690 of *Lecture Notes in Computer Science*, pages 244–254. Springer Berlin Heidelberg, 2009.

[16] E. F. Codd. A relational model of data for large shared data banks. *Commun. ACM*, 13(6):377–387, June 1970.

[17] Jeffrey Cohen, Brian Dolan, Mark Dunlap, Joseph M Hellerstein, and Caleb Welton. MAD skills: new analysis practices for big data. *Proceedings of the VLDB Endowment*, 2(2):1481–1492, 2009.

[18] Yann Collet. LZ4 - Extremely fast compression. Technical report, 2013.

[19] Andrew Crotty, Alex Galakatos, Kayhan Dursun, Tim Kraska, Carsten Binnig, Ugur Cetintemel, and Stan Zdonik. An architecture for compiling udf-centric workflows. *Proc. VLDB Endow.*, 8(12):1466–1477, August 2015.

[20] Anthony Damico. American Community Survey (ACS). Technical report, 2018.

[21] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[22] Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2019. R package version 1.12.0.

[23] Dwayne Richard Hipp. Create Or Redefine SQL Functions. In *SQLite User Manual*.

[24] Jon Ellis and Linda Ho. JDBC 3.0 Specification. Technical report, Sun Microsystems, October 2001.

[25] Greenplum Database 4.2. Technical report, EMC Corporation, 2012.

[26] Xixuan Feng, Arun Kumar, Benjamin Recht, and Christopher Ré. Towards a Unified Architecture for in-RDBMS Analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 325–336, New York, NY, USA, 2012. ACM.

[27] Lukas Fittl. C library for accessing the postgresql parser outside of the server environment. `https://github.com//fittl/libpg_query`, 2019.

[28] Free Software Foundation. The GNU C Library - Memory-mapped I/O. Technical report, Free Software Foundation, 2018.

[29] The GNU C Library - Memory Protection. Technical report, Free Software Foundation, 2018.

[30] Eric Friedman, Peter Pawlowski, and John Cieslewicz. Sql/mapreduce: A practical approach to self-describing, polymorphic, and parallelizable user-defined functions. *Proc. VLDB Endow.*, 2(2):1402–1413, August 2009.

[31] Jean Gailly. gzip: The data compression program. Technical report, University of Utah, July 1993.

[32] Kyle Geiger. *Inside ODBC*. Microsoft Press, 1995.

[33] Chris Gibson, Kris Katterjohn, Mixter, and Fyodor. Ncat Reference Guide. Technical report, Nmap project, 2016.

[34] Protocol Buffers: Developer's Guide. Technical report, Google, 2016.

[35] Anurag Gupta, Deepak Agarwal, Derek Tan, et al. Amazon Redshift and the Case for Simpler Data Warehouses. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 1917–1923, New York, NY, USA, 2015. ACM.

[36] Joseph M. Hellerstein, Christoper Ré, Florian Schoppmann, Daisy Zhe Wang, Eugene Fratkin, Aleksander Gorajek, Kee Siong Ng, Caleb Welton, Xixuan Feng,

Kun Li, and Arun Kumar. The MADlib Analytics Library: Or MAD Skills, the SQL. *Proc. VLDB Endow.*, 5(12):1700–1711, August 2012.

[37] Joseph M. Hellerstein, Christoper R, U. Wisconsin, Aleksander Gorajek, Kun Li, U. Florida, Kee Siong Ng, U. Wisconsin, Caleb Welton, Daisy Zhe Wang, U. Florida, Xixuan Feng, and U. Wisconsin. The MADlib analytics library, or MAD skills, the SQL.

[38] Joseph M. Hellerstein and Michael Stonebraker. Predicate migration: Optimizing queries with expensive predicates. *SIGMOD Rec.*, 22(2):267–276, June 1993.

[39] Stephen Hemminger. Network Emulation with NetEm. Technical report, Open Source Development Lab, April 2005.

[40] Pedro Holanda, Mark Raasveldt, and Martin Kersten. Don't Hold My UDFs Hostage - Exporting UDFs For Debugging Purposes. In *Proceedings of the 28th International Conference on Simpósio Brasileiro de Banco de Dados, SSBD 2017, Uberlndia, Brazil*, 2017.

[41] Stratos Idreos, Fabian Groffen, Niels Nes, Stefan Manegold, Sjoerd Mullender, and Martin Kersten. MonetDB: Two Decades of Research in Column-oriented Database Architectures. *IEEE Data Eng. Bull*, 2012.

[42] MongoDB Inc. MongoDB Architecture Guide. Technical report, MongoDB Inc., June 2016.

[43] ISO. Iso/iec 9075:1992, database language sql. Technical report, July 1992.

[44] Michael Jaedicke and Bernhard Mitschang. On parallel processing of aggregate and scalar functions in object-relational dbms. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, pages 379–389, New York, NY, USA, 1998. ACM.

[45] Michael Jaedicke and Bernhard Mitschang. User-defined table operators: Enhancing extensibility for ordbms. In Malcolm P. Atkinson, Maria E. Orlowska, Patrick

Valduriez, Stanley B. Zdonik, and Michael L. Brodie, editors, *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*, pages 494–505. Morgan Kaufmann, 1999.

[46] Steinar H. Gunderson Jeff Dean, Sanjay Ghemawat. Snappy, a fast compressor/decompressor. Technical report, Google, 2016.

[47] Roger Magoulas John King. 2015 data science salary survey. September 2015.

[48] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2015-10-01].

[49] Andrew Lamb, Matt Fuller, Ramakrishna Varadarajan, et al. The Vertica Analytic Database: C-store 7 Years Later. *Proc. VLDB Endow.*, 5(12):1790–1801, August 2012.

[50] Harald Lang, Tobias Mühlbauer, Florian Funke, et al. Data blocks: Hybrid OLTP and OLAP on compressed storage using both vectorization and compilation. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 311–326, 2016.

[51] Daniel Lemire and Leonid Boytsov. Decoding billions of integers per second through vectorization. *CoRR*, abs/1209.2137, 2012.

[52] Volker Linnemann, Klaus Küspert, Peter Dadam, Peter Pistor, R. Erbe, Alfons Kemper, Norbert Südkamp, Georg Walch, and Mechtild Wallrath. Design and implementation of an extensible database management system supporting user defined data types and functions. In *Proceedings of the 14th International Conference on Very Large Data Bases*, VLDB '88, pages 294–305, San Francisco, CA, USA, 1988. Morgan Kaufmann Publishers Inc.

[53] Jon Loeliger and Matthew McCullough. *Version Control with Git: Powerful tools and techniques for collaborative software development.* " O'Reilly Media, Inc.", 2012.

[54] Thomas Lumley. Package 'survey'. 2018.

[55] John C McCallum. Memory Prices (1957-2019). Technical report, April 2019.

[56] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[57] Guido Moerkotte and Thomas Neumann. Dynamic programming strikes back. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 539–552, 2008.

[58] Thomas Neumann. Efficiently Compiling Efficient Query Plans for Modern Hardware. *Proc. VLDB Endow.*, 4(9):539–550, June 2011.

[59] Thomas Neumann and Alfons Kemper. Unnesting arbitrary queries. In *Datenbanksysteme für Business, Technologie und Web (BTW), 16. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 4.-6.3.2015 in Hamburg, Germany. Proceedings*, pages 383–402, 2015.

[60] Thomas Neumann, Tobias Mühlbauer, and Alfons Kemper. Fast serializable multi-version concurrency control for main-memory database systems. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 677–689, 2015.

[61] Thomas Neumann and Bernhard Radke. Adaptive optimization of very large join queries. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 677–692, New York, NY, USA, 2018. ACM.

[62] Department of Transport Statistics. Airline On-Time Statistics and Delay Causes. Technical report, United States Department of Transportation, 2016.

[63] Carlos Ordonez and Sasi K Pitchaimalai. One-pass data mining algorithms in a DBMS with UDFs. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 1217–1220. ACM, 2011.

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[65] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[66] J. Postel. Transmission Control Protocol. RFC 793 (Standard), September 1981. Updated by RFCs 1122, 3168, 6093.

[67] PostgreSQL Development Team. Procedural Languages. In *PostgreSQL User Manual*.

[68] Andrew Prunicki. Apache Thrift. Technical report, Object Computing, Inc., June 2009.

[69] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[70] Mark Raasveldt and Hannes Mühleisen. Don't Hold My Data Hostage: A Case for Client Protocol Redesign. *Proc. VLDB Endow.*, 10(10):1022–1033, June 2017.

[71] Bert Rich. *Oracle Database Reference, 12c Release 1*. 2017.

[72] David Salomon. *Data Compression: The Complete Reference*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[73] Sunita Sarawagi, Shiby Thomas, and Rakesh Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, pages 343–354, New York, NY, USA, 1998. ACM.

[74] Felix Martin Schuhknecht, Jens Dittrich, and Ankur Sharma. RUMA Has It: Rewired User-space Memory Access is Possible! *Proc. VLDB Endow.*, 9(10):768–779, June 2016.

[75] Lefteris Sidirourgos and Martin Kersten. Column imprints: A secondary index structure. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 893–904, New York, NY, USA, 2013. ACM.

[76] Michael Stonebraker and Greg Kemnitz. The POSTGRES Next Generation Database Management System. *Commun. ACM*, 34(10):78–92, October 1991.

[77] Michael Stonebraker and Greg Kemnitz. The POSTGRES Next Generation Database Management System. *Commun. ACM*, 34(10):78–92, October 1991.

[78] Carl-Fredrik Sundlöf. In-database computations. Master's thesis, Royal Institute of Technology, Sweden, 10 2010.

[79] MySQL Development Team. Adding a New User-Defined Function. In *MySQL User Manual*.

[80] The HDF Group. Hierarchical Data Format, version 5. 1997-NNNN. http://www.hdfgroup.org/HDF5/.

[81] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, et al. Hive - a petabyte scale data warehouse using Hadoop. *2014 IEEE 30th International Conference on Data Engineering*, 0:996–1005, 2010.

[82] Transaction Processing Performance Council. TPC Benchmark H (Decision Support) Standard Specification. Technical report, Transaction Processing Performance Council, June 2013.

[83] Jan Urbaski. Postgres on the wire. In *PGCon 2016*, May 2014.

[84] S. van der Walt, S.C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, 13(2):22–30, March 2011.

[85] Manasi Vartak, Harihar Subramanyam, Wei-En Lee, Srinidhi Viswanathan, Saadiyah Husnoo, Samuel Madden, and Matei Zaharia. Model DB: a system

for machine learning model management. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, page 14. ACM, 2016.

[86] Haixun Wang and Carlo Zaniolo. User-defined aggregates in database languages. In Richard Connor and Alberto Mendelzon, editors, *Research Issues in Structured and Semistructured Database Programming*, volume 1949 of *Lecture Notes in Computer Science*, pages 43–60. Springer Berlin Heidelberg, 2000.

[87] Hadley Wickham. Package 'DBI'. 2018.

[88] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2018. R package version 0.7.8.

[89] Michael Widenius and Davis Axmark. *MySQL Reference Manual*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 1st edition, 2002.

[90] Florian Wolf, Iraklis Psaroudakis, Norman May, Anastasia Ailamaki, and Kai-Uwe Sattler. Extending database task schedulers for multi-threaded application code. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, SSDBM '15, pages 25:1–25:12, New York, NY, USA, 2015. ACM.

[91] Paul C. Zikopoulos and Roman B. Melnyk. *DB2: The Complete Reference*. McGraw-Hill Companies, January 2001.