《Architecture of a Database System》

（中文版）

Joseph M. Hellerstein, Michael Stonebraker and James Hamilton

翻译：林子雨

厦门大学数据库实验室

http://dblab.xmu.edu.cn

中文版网址：http://dblab.xmu.edu.cn/node/459

厦门大学计算机科学系教师 林子雨 翻译作品

http://www.cs.xmu.edu.cn/linziyu

2013 年 9 月

前言

本文翻译自经典英文论文《Architecture of a Database System》，原文作者是Joseph M. Hellerstein, Michael Stonebraker 和 James Hamilton。该论文可以作为中国各大高校数据库实验室研究生的入门读物，帮助学生快速了解数据库的内部运行机制。

本文一共包括 6 章，分别是：第 1 章概述，第 2 章进程模型，第 3 章并行体系结构：进程和内存协调，第 4 章关系查询处理器，第 5 章存储管理，第 6 章事务：并发控制和恢复，第 7 章共享组件，第 8 章结束语。

本文翻译由厦门大学数据库实验室林子雨老师团队合力完成，其中，林子雨老师负责统稿校对，刘颖杰同学负责翻译第 1 章、第 2 章和第 6 章，罗道文同学负责翻译第 3 章和第 4 章，谢荣东同学负责翻译第 5 章，蔡珉星同学负责翻译第 7 章和第 8 章，并负责对林子雨老师校对结果进行二次校对。

如果对本文翻译内容有任何疑问，欢迎联系林子雨老师。

林子雨的E-mail是：ziyulin@xmu.edu.cn。

林子雨的个人主页是：http://www.cs.xmu.edu.cn/linziyu。

厦门大学数据库实验室网站是：http://dblab.xmu.edu.cn。

本文中文版的网址是：http://dblab.xmu.edu.cn/node/459。

林子雨于厦门大学海韵园

2013 年 9 月

摘 要

数据库管理系统（DBMS）广泛存在于现代计算机系统中，并且是其重要的组成部分。它是学术界以及工业界数十年研究和发展的成果。在计算机发展史上，数据库属于最早开发的多用户服务系统之一，因此，它的研究也催生了许多为保证系统可拓展性以及稳定性的系统开发技术，这些技术如今被应用于许多其他的领域。虽然许多数据库的相关算法和概念广泛见于教科书中，但关于如何让一个数据库工作的系统设计问题却鲜有资料介绍。本文从体系架构角度探讨数据库设计的一些准则，包括处理模型、并行架构、存储系统设计、事务处理系统、查询处理及优化结构以及具有代表性的共享组件和应用。当业界有多种设计方式可供选择时，我们以当前成功的商业开源软件作为参考标准。

第 8 章 结束语

厦门大学计算机科学系教师 林子雨 编著

个人主页：http://www.cs.xmu.edu.cn/linziyu

中文版网址：http://dblab.xmu.edu.cn/node/459

2013 年 9 月

第 8 章 结束语

从本文中应该清楚的是，现代商业数据库系统是构建在两个基础之上的，一个是学术研究，另一个是为高端客户开发工业级别的产品所积累的大量经验。编写和维护一个高性能、全功能的关系型 DBMS 的任务，需要投入巨大的时间和精力。然而，许多关系型 DBMS 的经验已转化到新的领域。Web 服、网络附加存储、文本和电子邮件库、通知服务和网络监控等，这些领域都可以从 DBMS 的研究和经验中受益。数据密集型的服务是当今计算的核心，数据库系统设计的知识是可以被广泛应用到各个领域的技能，既包括数据库领域也包括其他领域。这些新的应用方向，也带来了一些数据库管理方面的研究问题，这为数据库社区和其他计算领域之间的交互开辟了新的道路。

致谢

英文原文的作者对以下人员表示感谢： Rob von Behren, Eric Brewer, Paul Brown, Amol Deshpande, Cesar Galindo-Legaria, Jim Gray, Wei Hong, Matt Huras, Lubor Kollar, Ganapathy Krishnamoorthy, Bruce Lindsay, Guy Lohman, S. Muralidhar, Pat Selinger, Mehul Shah 和 Matt。

参考文献

[1] A. Adya, B. Liskov, and P. O'Neil, "Generalized isolation level definitions," in 16th International Conference on Data Engineering (ICDE), San Diego, CA, February 2000.

[2] R. Agrawal, M. J. Carey, and M. Livny, "Concurrency control performance modelling: Alternatives and implications," ACM Transactions on Database Systems (TODS), vol. 12, pp. 609–654, 1987.

[3] M. M. Astrahan, M. W. Blasgen, D. D. Chamberlin, K. P. Eswaran, J. Gray, P. P. Griffiths, W. F. Frank King III, 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 Transactions on Database Systems (TODS), vol. 1, pp. 97–137, 1976.

[4] R. Bayer and M. Schkolnick, "Concurrency of operations on B-trees," Acta Informatica, vol. 9, pp. 1–21, 1977.

[5] K. P. Bennett, M. C. Ferris, and Y. E. Ioannidis, "A genetic algorithm for database query optimization," in Proceedings of the 4th International Conference on Genetic Algorithms, pp. 400–407, San Diego, CA, July 1991.

[6] H. Berenson, P. A. Bernstein, J. Gray, J. Melton, E. J. O'Neil, and P. E. O'Neil, "A critique of ANSI SQL isolation levels," in Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 1–10, San Jose, CA, May 1995.

[7] P. A. Bernstein and N. Goodman, "Concurrency control in distributed database systems," ACM Computing Surveys, vol. 13, 1981.

[8] W. Bridge, A. Joshi, M. Keihl, T. Lahiri, J. Loaiza, and N. MacNaughton, "The oracle universal server buffer," in Proceedings of 23rd International Conference on Very Large Data Bases (VLDB), pp. 590–594, Athens, Greece, August 1997.

[9] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra,

A. Fikes, and R. E. Gruber, “Bigtable: A distributed storage system for structured data,” in Symposium on Operating System Design and Implementation (OSDI), 2006.

[10] S. Chaudhuri, “An overview of query optimization in relational systems,” in Proceedings of ACM Principles of Database Systems (PODS), 1998.

[11] S. Chaudhuri and U. Dayal, “An overview of data warehousing and olap technology,” ACM SIGMOD Record, March 1997.

[12] S. Chaudhuri and V. R. Narasayya, “Autoadmin ‘what-if’ index analysis utility,” in Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 367–378, Seattle, WA, June 1998.

[13] S. Chaudhuri and K. Shim, “Optimization of queries with user-defined predicates,” ACM Transactions on Database Systems (TODS), vol. 24, pp. 177–228, 1999.

[14] M.-S. Chen, J. Hun, and P. S. Yu, “Data mining: An overview from a database perspective,” IEEE Transactions on Knowledge and Data Engineering, vol. 8, 1996.

[15] H.-T. Chou and D. J. DeWitt, “An evaluation of buffer management strategies for relational database systems,” in Proceedings of 11th International Conference on Very Large Data Bases (VLDB), pp. 127–141, Stockholm, Sweden, August 1985.

[16] A. Desphande, M. Garofalakis, and R. Rastogi, “Independence is good: Dependency-based histogram synopses for high-dimensional data,” in Proceedings of the 18th International Conference on Data Engineering, San Jose, CA, February 2001.

[17] P. Flajolet and G. Nigel Martin, “Probabilistic counting algorithms for data base applications,” Journal of Computing System Science, vol. 31, pp. 182–209, 1985.

[18] C. A. Galindo-Legaria, A. Pellenkoft, and M. L. Kersten, “Fast, randomized join-order selection — why use transformations?,” VLDB, pp. 85–95, 1994.

[19] S. Ganguly, W. Hasan, and R. Krishnamurthy, “Query optimization for parallel execution,” in Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 9–18, San Diego, CA, June 1992.

[20] M. Garofalakis and P. B. Gibbons, “Approximate query processing: Taming the terabytes, a tutorial,” in International Conferenence on Very Large Data Bases, 2001. www.vldb.org/conf/2001/tut4.pdf.

[21] M. N. Garofalakis and Y. E. Ioannidis, “Parallel query scheduling and optimization with

time- and space-shared resources," in Proceedings of 23rd International Conference on Very Large Data Bases (VLDB), pp. 296–305, Athens, Greece, August 1997.

[22] R. Goldman and J. Widom, "Wsq/dsq: A practical approach for combined querying of databases and the web," in Proceedings of ACM-SIGMOD International Conference on Management of Data, 2000.

[23] G. Graefe, "Encapsulation of parallelism in the volcano query processing system," in Proceedings of ACM-SIGMOD International Conference on Management of Data, pp. 102–111, Atlantic City, May 1990.

[24] G. Graefe, "Query evaluation techniques for large databases," Computing Surveys, vol. 25, pp. 73–170, 1993.

[25] G. Graefe, "The cascades framework for query optimization," IEEE Data Engineering Bulletin, vol. 18, pp. 19–29, 1995.

[26] C. Graham, "Market share: Relational database management systems by operating system, worldwide, 2005," Gartner Report No: G00141017, May 2006.

[27] J. Gray, "Greetings from a filesystem user," in Proceedings of the FAST '05 Conference on File and Storage Technologies, (San Francisco), December 2005.

[28] J. Gray and B. Fitzgerald, FLASH Disk Opportunity for Server-Applications. http://research.microsoft.com/～Gray/papers/FlashDiskPublic.doc.

[29] J. Gray, R. A. Lorie, G. R. Putzolu, and I. L. Traiger, "Granularity of locks and degrees of consistency in a shared data base," in IFIP Working Conference on Modelling in Data Base Management Systems, pp. 365–394, 1976.

[30] J. Gray and A. Reuter, Transaction Processing: Concepts and Techniques. Morgan Kaufmann, 1993.

[31] S. D. Gribble, E. A. Brewer, J. M. Hellerstein, and D. Culler, "Scalable, distributed data structures for internet service construction," in Proceedings of the Fourth Symposium on Operating Systems Design and Implementation (OSDI), 2000.

[32] A. Guttman, "R-trees: A dynamic index structure for spatial searching," in Proceedings of ACM-SIGMOD International Conference on Management of Data, pp. 47–57, Boston, June 1984.

[33] L. Haas, D. Kossmann, E. L. Wimmers, and J. Yang, "Optimizing queries across

diverse data sources," in International Conference on Very Large Databases (VLDB), 1997.

[34] T. Haerder and A. Reuter, "Principles of transaction-oriented database recovery," ACM Computing Surveys, vol. 15, pp. 287–317, 1983.

[35] S. Harizopoulos and N. Ailamaki, "StagedDB: Designing database servers for modern hardware," IEEE Data Engineering Bulletin, vol. 28, pp. 11–16, June 2005.

[36] S. Harizopoulos, V. Liang, D. Abadi, and S. Madden, "Performance tradeoffs in read-optimized databases," in Proceedings of the 32nd Very Large Databases Conference (VLDB), 2006.

[37] J. M. Hellerstein, "Optimization techniques for queries with expensive methods," ACM Transactions on Database Systems (TODS), vol. 23, pp. 113–157, 1998.

[38] J. M. Hellerstein, P. J. Haas, and H. J. Wang, "Online aggregation," in Proceedings of ACM-SIGMOD International Conference on Management of Data, 1997.

[39] J. M. Hellerstein, J. Naughton, and A. Pfeffer, "Generalized search trees for database system," in Proceedings of Very Large Data Bases Conference (VLDB), 1995.

[40] J. M. Hellerstein and A. Pfeffer, "The russian-doll tree, an index structure for sets," University of Wisconsin Technical Report TR1252, 1994.

[41] C. Hoare, "Monitors: An operating system structuring concept," Communications of the ACM (CACM), vol. 17, pp. 549–557, 1974.

[42] W. Hong and M. Stonebraker, "Optimization of parallel query execution plans in xprs," in Proceedings of the First International Conference on Parallel and Distributed Information Systems (PDIS), pp. 218–225, Miami Beach, FL, December 1991.

[43] H.-I. Hsiao and D. J. DeWitt, "Chained declustering: A new availability strategy for multiprocessor database machines," in Proceedings of Sixth International Conference on Data Engineering (ICDE), pp. 456–465, Los Angeles, CA, November 1990.

[44] Y. E. Ioannidis and Y. Cha Kang, "Randomized algorithms for optimizing large join queries," in Proceedings of ACM-SIGMOD International Conference on Management of Data, pp. 312–321, Atlantic City, May 1990.

[45] Y. E. Ioannidis and S. Christodoulakis, "On the propagation of errors in the size of join results," in Proceedings of the ACM SIGMOD International Conference on Management

of Data, pp. 268–277, Denver, CO, May 1991.

[46] M. Kornacker, C. Mohan, and J. M. Hellerstein, “Concurrency and recovery in generalized search trees,” in Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 62–72, Tucson, AZ, May 1997.

[47] H. T. Kung and J. T. Robinson, “On optimistic methods for concurrency control,” ACM Tranactions on Database Systems (TODS), vol. 6, pp. 213–226, 1981.

[48] J. R. Larus and M. Parkes, “Using cohort scheduling to enhance server performance,” in USENIX Annual Conference, 2002.

[49] H. C. Lauer and R. M. Needham, “On the duality of operating system structures,” ACM SIGOPS Operating Systems Review, vol. 13, pp. 3–19, April 1979.

[50] P. L. Lehman and S. Bing Yao, “Efficient locking for concurrent operations on b-trees,” ACM Transactions on Database Systems (TODS), vol. 6, pp. 650–670, December 1981.

[51] A. Y. Levy, “Answering queries using views,” VLDB Journal, vol. 10, pp. 270–294, 2001.

[52] A. Y. Levy, I. Singh Mumick, and Y. Sagiv, “Query optimization by predicate move-around,” in Proceedings of 20th International Conference on Very Large Data Bases, pp. 96–107, Santiago, September 1994.

[53] W. Litwin, “Linear hashing: A new tool for file and table addressing,” in Sixth International Conference on Very Large Data Bases (VLDB), pp. 212–223, Montreal, Quebec, Canada, October 1980.

[54] G. M. Lohman, “Grammar-like functional rules for representing query optimization alternatives,” in Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 18–27, Chicago, IL, June 1988.

[55] Q. Luo, S. Krishnamurthy, C. Mohan, H. Pirahesh, H. Woo, B. G. Lindsay, and J. F. Naughton, “Middle-tier database caching for e-business,” in Proceedings of ACM SIGMOD International Conference on Management of Data, 2002.

[56] S. R. Madden and M. J. Franklin, “Fjording the stream: An architecture for queries over streaming sensor data,” in Proceedings of 12th IEEE International Conference on Data Engineering (ICDE), San Jose, February 2002.

[57] V. Markl, G. Lohman, and V. Raman, “Leo: An autonomic query optimizer for db2,”

IBM Systems Journal, vol. 42, pp. 98–106, 2003.

[58] C. Mohan, “Aries/kvl: A key-value locking method for concurrency control of multiaction transactions operating on b-tree indexes,” in 16th International Conference on Very Large Data Bases (VLDB), pp. 392–405, Brisbane, Queensland, Australia, August 1990.

[59] C. Mohan, D. J. Haderle, B. G. Lindsay, H. Pirahesh, and P. M. Schwarz, “Aries: A transaction recovery method supporting fine-granularity locking and partial rollbacks using write-ahead logging,” ACM Transactions on Database Systems (TODS), vol. 17, pp. 94–162, 1992.

[60] C. Mohan and F. Levine, “Aries/im: An efficient and high concurrency index management method using write-ahead logging,” in Proceedings of ACM SIGMOD International Conference on Management of Data, (M. Stonebraker, ed.), pp. 371–380, San Diego, CA, June 1992.

[61] C. Mohan, B. G. Lindsay, and R. Obermarck, “Transaction management in the r* distributed database management system,” ACM Transactions on Database Systems (TODS), vol. 11, pp. 378–396, 1986.

[62] E. Nightingale, K. Veerarghavan, P. M. Chen, and J. Flinn, “Rethink the sync,” in Symposium on Operating Systems Design and Implementation (OSDI), November 2006.

[63] OLAP Market Report. Online manuscript. http://www.olapreport.com/market.htm.

[64] E. J. O’Neil, P. E. O’Neil, and G. Weikum, “The lru-k page replacement algorithm for database disk buffering,” in Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 297–306, Washington, DC, May 1993.

[65] P. E. O’Neil and D. Quass, “Improved query performance with variant indexes,” in Proceedings of ACM-SIGMOD International Conference on Management of Data, pp. 38–49, Tucson, May 1997.

[66] S. Padmanabhan, B. Bhattacharjee, T. Malkemus, L. Cranston, and M. Huras, “Multi-dimensional clustering: A new data layout scheme in db2,” in ACM SIGMOD International Management of Data (San Diego, California, June 09–12, 2003) SIGMOD ’03, pp. 637–641, New York, NY: ACM Press, 2003.

[67] D. Patterson, “Latency lags bandwidth,” CACM, vol. 47, pp. 71–75, October 2004.

[68] H. Pirahesh, J. M. Hellerstein, and W. Hasan, “Extensible/rule- based query rewrite optimization in starburst,” in Proceedings of ACM-SIGMOD International Conference on Management of Data, pp. 39–48, San Diego, June 1992.

[69] V. Poosala and Y. E. Ioannidis, “Selectivity estimation without the attribute value independence assumption,” in Proceedings of 23rd International Conference on Very Large Data Bases (VLDB), pp. 486–495, Athens, Greece, August 1997.

[70] M. P¨oss, B. Smith, L. Koll´ar, and P.-˚A. Larson, “Tpc-ds, taking decision support benchmarking to the next level,” in SIGMOD 2002, pp. 582–587.

[71] V. Prabakharan, A. C. Arpaci-Dusseau, and R. Arpaci-Dusseau, “Analysis and evolution of journaling file systems,” in Proceedings of USENIX Annual Technical Conference, April 2005.

[72] R. Ramakrishnan and J. Gehrke, “Database management systems,” McGraw-Hill, Boston, MA, Third ed., 2003.

[73] V. Raman and G. Swart, “How to wring a table dry: Entropy compression of relations and querying of compressed relations,” in Proceedings of International Conference on Very Large Data Bases (VLDB), 2006.

[74] D. P. Reed, Naming and Synchronization in a Decentralized Computer System. PhD thesis, MIT, Dept. of Electrical Engineering, 1978.

[75] A. Reiter, “A study of buffer management policies for data management systems,” Technical Summary Report 1619, Mathematics Research Center, University of Wisconsin, Madison, 1976.

[76] D. J. Rosenkrantz, R. E. Stearns, and P. M. Lewis, “System level concurrency control for distributed database systems,” ACM Transactions on Database Systems (TODS), vol. 3, pp. 178–198, June 1978.

[77] S. Sarawagi, S. Thomas, and R. Agrawal, “Integrating mining with relational database systems: Alternatives and implications,” in Proceedings of ACMSIGMOD International Conference on Management of Data, 1998.

[78] R. Sears and E. Brewer, “Statis: Flexible transactional storage,” in Proceedings of Symposium on Operating Systems Design and Implementation (OSDI), 2006.

[79] P. G. Selinger, M. Astrahan, D. Chamberlin, R. Lorie, and T. Price, “Access path

selection in a relational database management system," in Proceedings of ACM-SIGMOD International Conference on Management of Data, pp. 22–34, Boston, June 1979.

[80] P. Seshadri, H. Pirahesh, and T. Y. C. Leung, "Complex query decorrelation," in Proceedings of 12th IEEE International Conference on Data Engineering (ICDE), New Orleans, February 1996.

[81] M. A. Shah, S. Madden, M. J. Franklin, and J. M. Hellerstein, "Java support for data-intensive systems: Experiences building the telegraph dataflow system," ACM SIGMOD Record, vol. 30, pp. 103–114, 2001.

[82] L. D. Shapiro, "Exploiting upper and lower bounds in top-down query optimization," International Database Engineering and Application Symposium (IDEAS), 2001.

[83] A. Silberschatz, H. F. Korth, and S. Sudarshan, Database System Concepts. McGraw-Hill, Boston, MA, Fourth ed., 2001.

[84] M. Steinbrunn, G. Moerkotte, and A. Kemper, "Heuristic and randomized optimization for the join ordering problem," VLDB Journal, vol. 6, pp. 191–208, 1997.

[85] M. Stonebraker, "Retrospection on a database system," ACM Transactions on Database Systems (TODS), vol. 5, pp. 225–240, 1980.

[86] M. Stonebraker, "Operating system support for database management," Communications of the ACM (CACM), vol. 24, pp. 412–418, 1981.

[87] M. Stonebraker, "The case for shared nothing," IEEE Database Engineering Bulletin, vol. 9, pp. 4–9, 1986.

[88] M. Stonebraker, "Inclusion of new types in relational data base systems," ICDE, pp. 262–269, 1986.

[89] M. Stonebraker, D. J. Abadi, A. Batkin, X. Chen, M. Cherniack, M. Ferreira, E. Lau, A. Lin, S. Madden, E. O'Neil, P. O'Neil, A. Rasin, N. Tran, and S. Zdonik, "C-store: A column oriented dbms," in Proceedings of the Conference on Very Large Databases (VLDB), 2005.

[90] M. Stonebraker and U. Cetintemel, "One size fits all: An idea whose time has come and gone," in Proceedings of the International Conference on Data Engineering (ICDE), 2005.

[91] Transaction Processing Performance Council 2006. TPC Benchmark C Standard

Specification Revision 5.7, http://www.tpc.org/tpcc/spec/tpcc current. pdf, April.

[92] T. Urhan, M. J. Franklin, and L. Amsaleg, “Cost based query scrambling for initial delays,” ACM-SIGMOD International Conference on Management of Data, 1998.

[93] R. von Behren, J. Condit, F. Zhou, G. C. Necula, and E. Brewer, “Capriccio: Scalable threads for internet services,” in Proceedings of the Ninteenth Symposium on Operating System Principles (SOSP-19), Lake George, New York, October 2003.

[94] M. Welsh, D. Culler, and E. Brewer, “Seda: An architecture for well- conditioned, scalable internet services,” in Proceedings of the 18th Symposium on Operating Systems Principles (SOSP-18), Banff, Canada, October 2001.

[95] C. Zou and B. Salzberg, “On-line reorganization of sparsely-populated b+trees,” pp. 115–124, 1996.

附录 1:译者介绍

林子雨(1978－),男,博士,厦门大学计算机科学系助理教授,主要研究领域为数据库,数据仓库,数据挖掘.
主讲课程：《大数据技术基础》
办公地点：厦门大学海韵园科研 2 号楼
E-mail: ziyulin@xmu.edu.cn
个人网页：http://www.cs.xmu.edu.cn/linziyu