Database Qualifying Exam Reading List

The Database Management Systems (DBMS) qualifying exam is intended to cover a wide range of database systems literature, testing the students' preparedness and ability for pursuing serious research in the database area. As such, students who do not intend to pursue database-related Ph.D. topics are advised not to attempt taking this exam. The exam will be graded with the expectation that each examinee will have a fairly deep understanding of the basic issues plus a broad knowledge of recent database research.

The exam is intended to cover a set of fundamental topics in databases. Students should have a solid grasp of the database systems area at the advanced undergraduate/introductory graduate level, as can be obtained by studying Ramakrishnan and Gerhke, or a comparable textbook, and in addition, they should have a deep understanding of many of the advanced topics discussed in the research literature. The following list of topics and papers indicates the material that all students will be expected to be familiar with. Students are encouraged to take the following (or equivalent) courses in preparation for the database qualifying exam.

- CS 564 *Database Management Systems: Design and Implementation*
- CS 764 *Topics in Database Management Systems*
- CS 784 *Foundations of Data Management*
- CS 774 (previous CS 784)

References

You are expected to be familiar with the coverage of the following topics in the text below (or a comparable text book): file organizations, indexing (B-tree and hash), database design, data models and languages, data mining, decision support, data warehousing, deductive databases, information retrieval, object-oriented and object-relational databases, query processing, transaction management, view maintenance, security and integrity, and XML.

- Ramakrishnan, R., and Gehrke, J., *Database Management Systems*, 3rd Edition, McGraw-Hill, 2003.

Further required reading, in the form of papers and manuscripts from the research literature, is listed below.

Data Models and Languages

- Codd, E. F., *A Relational Model of Data for Large Shared Data Banks*, Communications of the ACM, 1970.
- J. Ullman. *Database and Knowledge Base Systems*, vol. I. Chapter 3 (Logic as a Data Model).
- Stonebraker, M., *Inclusion of New Types In Relational Data Base Systems*, Proceedings of the International Conference on Data Engineering, 1986.
- Stonebraker M., and Hellerstein J., *What Goes Around Comes Around.*

Database Theory

- S. Abiteboul, R. Hull, V. Vianu. *Foundations of Databases*. Available for free from: `http://webdam.inria.fr/Alice/`
  - Conjunctive Queries (Chapters 3,4)
  - Chapter 6, Sections 6.2 and 6.4
  - Datalog (Chapter 12, Sections 12.1 - 12.3, Chapter 13, Section 13.1 - 13.3)
- T. J. Green, S. Huang, B. T. Loo, W. Zhou, *Datalog and Recursive Query Processing*, Foundations and Trends in Databases, Vol. 5, 2012.
- H. Ngo, C. Re, A. Rudra, *Skew Strikes Back: New Developments in the Theory of Join Algorithms*, SIGMOD Record, 2013.
- Cheney, Chiticariu, Tan, *Provenance in Databases: Why, How and Where*, Foundations and Trends in Databases, 2009.
- C. Dwork, *A Firm Foundation for Private Data Analysis*, Communications of the ACM, 2011.

DBMS Architecture

- Chamberlin, D. D., Astrahan, M. M., Blasgen, M. W., Gray, J. N., King, W. F., Lindsay, B. G., Lorie, R., Mehl, J. W., Price, T. G., Putzolu, F., Selinger, P. G., Schkolnick, M., Slutz, D. R., Traiger, I. L., Wade, B. W. and Yost, R. A. *A History and Evaluation of System R*, Communications of the ACM 24(10), 1981.
- Stonebraker, M., Wong E., Kreps P., and Held G., *The Design and Implementation of INGRES*, ACM Transactions on Database Systems 1(3), 1976.
- Hellerstein J. M., Stonebraker M. and Hamilton, J. R., *Architecture of a Database System*, Foundations and Trends in Databases 1(2), 2007.

Operating System Issues

- Chou, H., and DeWitt, D., *An Evaluation of Buffer Management Strategies for Relational Database Systems*, Proceedings of the International Conference on Very Large Data Bases (VLDB), 1985.
- O'Neil, E. J., O'Neil, P. E., Weikum G., *The LRU-K Page Replacement Algorithm For Database Disk Buffering*, Proceedings of the ACM-SIGMOD International Conference on Management of Data, 1993.
- Stonebraker, M., *Operating System Support for Database Management*, Communications of the ACM 24(7), 1981.

File Organizations and Access Methods

- Comer, D., *The Ubiquitous B-Tree*, ACM Computing Surveys 11(2), June 1979.
- Guttman, A., *R-Trees: A Dynamic Index Structure for Spatial Searching*, SIGMOD Conference, 1984.
- Patrick E. O'Neil, Dallan Quass: *Improved Query Performance with Variant Indexes*. SIGMOD Conference, 1997: 38-49.

Query Processing and Optimization

- Shapiro, L. D., *Join Processing in Database Systems with Large Main Memories*, ACM Transactions on Database Systems 11(3), 1986.
- Selinger, P., et al., *Access Path Selection in a Relational Database Management System*, Proceedings of the ACM-SIGMOD International Conference on Management of Data, 1979.
- Surajit Chaudhuri: *An Overview of Query Optimization in Relational Systems.* PODS 1998: 34-43
- Goetz Graefe: *Query Evaluation Techniques for Large Databases.* ACM Comput. Surv. 25(2): 73-170 (1993)

Concurrency Control and Recovery

- Bernstein, P.A., Hadzilacos, V., and Goodman, N., *Concurrency Control and Recovery in Database Systems*, Addison-Wesley, 1987; can be freely downloaded from Bernstein's webpage. (Chapters 1 and 2)
- Gray, J., Lorie, R. A., Pulzolu, G. R., Traiger, I. L., *Granularity of Locks and Degrees of Consistency in a Shared Data Base*, Proceedings of the IFIP Working Conference on Modeling of Data Base Management Systems, 1979.
- Kung, H., and Robinson, J., *On Optimistic Methods for Concurrency Control*, ACM Transactions on Database Systems 6(2), June 1981.
- Berenson, H., Bernstein, P. A., Gray, J., Melton, J., O'Neil, E. J., O'Neil, P. E., *A Critique of ANSI SQL Isolation Levels*, Proceedings of the ACM-SIGMOD International Conference on Management of Data, 1995.
- Lehman, P. and Yao, S., *Efficient Locking for Concurrent Operations on B-Trees*, ACM Transactions on Database Systems, 6(4): 650-670.
- Mohan, C., Haderle, D., Lindsay, B., Pirahesh, H., Schwarz, P., *ARIES: A Transaction Recovery Method Supporting Fine-Granularity Locking and Partial Rollbacks Using Write-Ahead Logging*, ACM Transactions on Database Systems, 17(1): 94-162.

Distributed and Parallel Data Processing

- Donald Kossmann, *The state of the art in distributed query processing.* ACM Comput. Surv. 32, 4 December 2000.
- Lohman, G., and Mackert, L., *R* Optimizer Validation and Performance Evaluation for Distributed Queries*, Proceedings of the 12th International Conference on Very Large Data Bases, 1986.
- Mohan, C., Lindsay, B., and Obermarck R., *Transaction Management in the R* Distributed Database Management System*, ACM Transactions on Database Systems 11 (4), 1986.
- DeWitt, D., et al, *The GAMMA Database Machine Project*, IEEE Transactions on Knowledge and Data Engineering 2(1), 1990.
- Dean J. and Ghemawat S., *MapReduce: Simplified Data Processing on Large Clusters*, Proceedings of the Symposium on Operating System Design and Implementation (OSDI) 2004.

Data Analysis and Decision Support

- Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart D., Venkatrao, M., Pellow, F., and Pirahesh, H., *Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub- Totals.* In Data Mining and Knowledge Discovery, 1(1): 29?53.
- Agrawal, R., and R. Srikant, *Fast Algorithms for Mining Association Rules.* In Proceedings of the 20th International Conference on Very Large Data Bases, 487?499.
- Zhang, T., Ramakrishnan R., and Livny M., *BIRCH: A Clustering Algorithm for Large Multidimensional Datasets*, Proceedings of the ACM SIGMOD International Conference on Management of Data, 1996
- Graham Cormode, Minos Garofalakis, Peter J. Haas and Chris Jermaine, *Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches*, Foundations and Trends in Databases, Vol. 4, 2011.

Data Extraction and Integration

- S. Sarawagi, *Information Extraction*, Foundations and Trends in Databases. Vol. 1, No. 3 (2007): read only Chapters 1-3.
- Levy, Alon, *Logic-based Techniques in Data Integration*, available on the Web at `http://homes.cs.washington.edu/~alon/site/files/levy-di00.ps`: read up to and including Section 5.1.
- Doan, A., Halevy, A., Ives, Z., *Principles of Data Integration*, Chapter 1, Chapter 5, Chapter 4: read 4.1, 4.2.1 (only Edit Distance), 4.2.2 (only Overlap, Jaccard, and TF/IDF), 4.2.4, and 4.3 (only Inverted Index and Size Filtering). Chapter 7: up to and including 7.5.3. Chapter 9: read 9.1, 9.2, and 9.3.1. Chapters available from `http://pages.cs.wisc.edu/~anhai/courses/dibook-chapters`.

DBMS and Search Engines

- Brin, S. and Page, L., *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Proceedings of Computer Networks and ISDN Systems, 1998.
- Singhal, A., *Modern Information Retrieval: a Brief Overview.* IEEE Data Engineering Bulletin, 24(4), 35- 43, 2001.
- Page, L. and Brin, S. and Motwani, R. and Winograd, T., *The Pagerank Citation Ranking: Bringing Order to the Web*, Technical Report, 1999.

Emerging Topics

- Jagadish et al., *Big Data and Its Technical Challenges*, CACM, Vol. 57 No. 7, 2014.
- Abadi et al., *The Beckman Report on Database Research*, SIGMOD Record 43(3), 2014.