Information Packet

CSSE 433

Advanced Database Systems
Spring 2008

Computer Science and Software Engineering
Rose-Hulman Institute of Technology

Computer Science and Software Engineering 433
Advanced Database Systems
Spring 2008

Instructor: Sriram Mohan

Office: Moench Hall, Room F203

Phones: (work) 877-8819; (cell) 812-219-9658

Email: mohan@rose-hulman.edu

Teaching Assistants:
Mark Jenne

Office Hours: I am usually in my office. Just stop by when you have questions.

Course Prerequisite: CSSE 333 **(**Introduction to Database Systems) or equivalent

Course Description: CSSE 433 covers advanced topics in the design and development of database management systems and their modern applications. Topics to be covered include

1. XML Databases
2. Data Security
3. Data Mining
4. Query Processing
5. Query Optimization Strategies
6. Transaction Management
7. Concurrency Control

We will also cover other trends in databases including Data Warehousing, Semi-structured Data Management, Data integration and Object Oriented Databases.

Course Outcomes: Students who complete this course will be able to

1) Write queries in Relational Calculus and Relational Algebra
2) Explain the basic concepts of the XML data model and specify queries in XPath and XQuery
3) Explain the basics behind query optimization and query evaluation techniques
4) Explain basic storage structures and access techniques for relational and post-relational systems
5) Explain the basics of data security and access control for relational and post-relational systems
6) Discuss the latest trends and research in database systems
7) Develop proof of concepts systems in databases that demonstrate the ability to perform research in databases

Grading:

The learning objectives for this course are listed above. We will be using various mechanisms to measure your progress against these learning objectives. The relative weights for each of these mechanisms are mentioned below:

Mechanism	Weight
Participation	10%
Exams	25%
Homework	10%
Paper Summaries	15%
Project	40%

Course Grade Division:

Metrics used to assign final grades are mentioned below. Please note that these are subject to change and the class will be notified accordingly.

90-100	A
85-89	B+
80-84	B
75-79	C+
70-74	C
65-69	D+
60-64	D
0-59	F

Exam Policy:

Exams will be in-class, closed book, and closed notes except for one 8.5 by 11 sheet of paper which you can put notes on using both sides of the page. No exams will be "dropped". If you have a conflict with a scheduled exam, you should notify me immediately. Giving a makeup exam for an unexcused absence is at the discretion of the instructor. Any requests for re-grading must be made in writing by the beginning of the next class period after the exams are returned.

Homework Grade:

There will be approximately 4 homework assignments. The homework's include a variety of tasks that will help each student understand the material covered in class.

Reading Assignment/Paper Summaries:

Most lectures will have an associated reading assignment. The reading assignment has to be completed before class on the day indicated. Refer to the course calendar for details on Reading Assignments.

On average, 3-4 papers/chapters will be discussed each week. All students are expected to read the papers/chapters BEFORE class. Students are required to submit a brief summary based on their reading before class. The paper review should summarize the key contribution and limitation of the paper. You may discuss the paper and the problem with your classmates. The write-up (summary or solution to the problem) should be written solely in your own words.

After the in-class discussion, if your opinion on the paper has changed, update your review and turn it to me along with the original review by the end of the day.

Ethics and Professional Practice:

You are expected to act honestly and professionally in this course at all times, in a manner consistent with the school's honor code.

Class Participation Policy:

There are 40 meeting times during the term. You can potentially receive 10 points towards the class participation portion of your grade for each of those classes in the following fashion:

- If there is a quiz during class, you can earn up to 10 points on it.
- If there is no quiz during class and you attend and make an effort to participate (with a small class there will be lots of discussion), you will earn 10 points.
- If the class for some reason does not meet, you automatically receive 10 points.

Late Submissions:

Please note that homework will be due at 11:55 PM, Friday of the week indicated. Paper Summaries are due in class on the day indicated. Project Deliverables will be due at 11:55 PM on the day indicated.

Late quizzes and paper summaries will not be accepted. Homework assignments, and Project milestone deliverables will also not be accepted late, with the following exception:

You have four "late day" credits. You may use one of them on any Homework assignment, or Project deliverable, which will allow you to submit that assignment up to 48 hours after the due time. Homework's or project assignments, which are more than 48 hours late, will receive a deduction of at least 10% per late day (or not be accepted at all), depending on the circumstances and the degree of lateness.

You may earn a maximum of one additional "late day" by submitting an assignment or a project deliverable 24 hours before the due date. Please send me an email alerting me to the same to obtain the "late day" credit.

If you submit something late for which late day credits are allowed, I will assume that you want to use one of your credits unless you tell me otherwise.

General Writing Issues:

Written communication is an important component of your final project, as it is in the profession in general. Remember that a document has several unique and important characteristics:

1. Technical documents are often the result of group authorship, thus it requires planning and final tweaking.
2. Specificity and organization are more important than flow, hence technical documentation is often ordered around lists and tables rather than paragraphs.
3. Documentation is often the reader's only source of information on the particular subject or product, hence it must be thorough and complete.
4. Documentation is often used to answer a specific question, hence it should facilitate finding a specific piece of information (navigation).
5. Documentation must bridge from general specifications to particulars of implementation and operation, hence it must make abstract concepts concrete and make concrete facts fit generalized concepts.
6. Documentation can be presented in many forms: online via HTML, MS help files, just plain text, and on paper as reference manuals, tutorial, quick reference guides, etc. It is important to choose the correct medium and even more important to write to fit the medium.

You can always drop by my office, if you have any questions regarding your document. I would be happy to look at it and suggest some changes. You should also be aware of the service provided by the Learning Center.

Attendance Policy:

Up to 2 unexcused absences allowed. In accordance with the Rose-Hulman attendance policy, additional unexcused absences may result in you receiving a failing grade for the course. You are responsible for making up any missed work.

Laptop Policy:

During class discussion, please do not use your laptops. Laptop use during discussions is distracting to your classmates and also keeps you from focusing on the material. If you typically use your laptop for note taking, please talk to your instructor so he can make an exception.

Collaboration:

You are encouraged to discuss the homework and other parts of the class with other students. Such discussions about ideas are not cheating, whereas the exchange of code or written answers is cheating. However, in such discussions of ideas, you should distinguish between helping and hurting yourself and the other student. In brief, you can help the other student by teaching them, and you can hurt them by giving them answers that they should

have worked out for themselves. The same applies to tutoring and getting help from the instructor.

Final Project

Team Composition:

Teams will consist of about 3-4 members each. We are willing to allow smaller teams, if you have a good reason for wanting to do so. We may consider allowing larger teams, but would require a very strong argument and a very interesting project idea before allowing this. Because a significant amount of class time will be devoted to project work during the last third of the term, we require that all members of a team attend the same section.

Project Ideas:

Students are recommended to choose from the topics suggested by the instructor, but they can also come up with their own project ideas, the following are the basic requirements all projects have to meet:

* Utilize the knowledge and skills learned in this course.
* It should be a research project, not a database application.
* Students are required to approach the project in the way they approach a research project, following the proper procedures including, but not limited to literature review, design, implementation and testing. Various documents and presentation are to be delivered at each milestone checkpoint. Here are a few projects ideas for student teams to consider:

1. Relational Algebraic Evaluator: This will be a pedagogical tool aimed at helping students understand the use of Relational algebra. The objective of the project will be to develop a tool that can be integrated with a Relational Database Management System to evaluate Relational algebraic expressions.

2. XPath Evaluator: This will be a pedagogical tool aimed at helping students understand the use of XPath. The objective of the project will be to develop a stand tool that can be integrated with a web browser to evaluate XPath Expressions.

3. SQL to Relational Algebra Converter: This will be a pedagogical tool aimed at helping students understand query optimization. The objective of the project will be to develop a tool that can generate multiple relational algebraic expressions for an SQL expression and demonstrate the benefits of query optimization.

4. RDBMS – XML Export Engine: The objective of this project will be to develop a tool that con export data and meta-data available in relational databases to semi-structured databases such as XML.

5. Access Control for Semi-structured Data: There has been a steady increase in the amount of data represented in semi-structured databases. The objective of this

project is to develop a technique/tool that can be used to specify and enforce access control on semi-structured databases.

6. Inference Control: An important area of research in data security deals with the notion of inference. Anti-inference deals with the study of information leakage and devising techniques to prevent such leakage. The objective of this project will be to develop an algorithm that can be used to detect instances of information leakage based on functional and other known dependencies.

7. Facebook Radial Searching: The objective of this project will be to analyze develop search algorithms that can used to perform efficient weighted searches on graph structures.

8. Transaction Management: This will be a pedagogical tool to help students understand the notion of transactions. The objective of the project will be to develop a simulation engine that can mimic the functioning of a transaction log for a simple text editor.

9. Access Control Migration: There is frequent migration of data between relational database systems and semi-structured files. How ever, when the data is migrated from relational systems, the access control rights have to be manually configured on the semi-structured data files and vice-versa. The objective of the project will be to develop a tool/technique that can automate the above process.

Milestones:

At each milestone, each team is required to submit the required documents. All team members are required to participate in generating and writing the milestone documents. A short presentation is also required at each milestone. (Slides of the presentation should be submitted as part of the milestone or final document.)

Milestone 1: Project proposal and literature review (Due Week 5 Day 1)

During this phase, the teams and the project topics should be finalized and the instructor notified of the same. Each team will research and understand the objectives and the goal of the project and completely a survey of related papers. At the end of this phase, each team will produce a project proposal that include 2-page project description, a short team and project organization plan, and an initial list of papers of reference. To be more specific, the project proposal should include:

- Project goal: what the project is about and what's going to be accomplished.
- Initial list of referenced papers
- Group organization (members, task/role)
- Proposed timeline

Milestone 2: Literature review and design (Due Week 6 Day 4)

Students will read the referenced papers, and conduct research in this stage. Depending on the nature of the project, the research may be to design a new algorithm or to improve an existing algorithm, or to collect data.

At the end of this stage, a milestone report will be produced, which should include an analysis of the state of art in the field (related to the project), and a summary of the research result up to the point.

Intermediate Status Reports:

Each team has complete freedom in deciding their internal organization and workload arrangement. Each team is expected to meet with the instructor on a weekly basis, to discuss the project progress. To ensure that the teams are making adequate progress, each team will demonstrate the current status of the project at the end of the 8th week.

Milestone 3: Project delivery: Demo and Presentation (Due Week 10 Day 4)

Each team will spend the rest of the quarter completing their project. Each team will demonstrate their final project and present it to the class during 10th week.

A final report; a research paper edited using the ACM (Association for Computing Machinery) SIGMOD (Special Interest Group in the Management of Data) format (double column, 10-12 pages[1]) and the code should be submitted before the presentation. The report should include the all components in previous reports (of course, refined), and the implementation details and results.

Each team member must independently complete a form evaluating his or her contribution to the team, and the contributions of the other team members. We will provide this form.

Project Grading

Grade will be determined based on several aspects. The quality of the research, the milestone reports and the final demonstration will determine the major portion of the project grade. Creativity and extensibility in design is highly appreciated. Other issues, such as teamwork will also be taken into consideration.

We will not necessarily assign the same grade to each member of a team. We reserve the right to adjust individual grades up or down based on peer evaluations and our observations about your teamwork. A team member can hurt his or her team either by being a slacker or by running roughshod over other team members. Thus, taking over a project and doing all the work might actually hurt your grade.

Papers:

[1] See http://www.acm.org/sigs/publications/proceedings-templates

The following is a list of papers that we will read during class this quarter. The papers have been organized by topic. You can download the papers from Angel.

Introduction To Relational Databases and XML (Weeks 1 and 2)

1. [Codd70] A Relational Model of Data for Large Shared Data Banks. E.F. Codd, CACM 13(6), 1970, pp 377-387.

2. [Astr76] System R: Relational Approach to DB Mgmt. Astrahan et al. ACM Trans. Database Systems 1(2), 1976, pp. 97-137.

3. [Ston76] The Design and Implementation of INGRES. Stonebraker et al, ACM Trans. Database Systems 1(3), 1976, pp. 189-222.

4. [XML] XML for the Absolute Beginner. Java World.[2]

5. [XMLDB] XML and Databases. Ronald Bourret.[3]

(Life Cycle of a Query/Spatial Databases) Week 3, 4 and 5

6. [Chou85] An Evaluation of Buffer Management Strategies for RDBMS. Hong-Tai Chou et al. VLDB 1985, pp 127-141

7. [Gaed98] Multidimensional Access Methods. Volker Gaede et al. ACM Computing Surveys, Vol. 30, No. 2, June 1998.

8. [Hell95] Generalized Search Trees for Database Systems. Joseph M. Hellerstein et al. VLDB 1995. pp 562-573.

9. [Gutt84] R-TREES. A Dynamic Index Structure For Spatial Searching. Antomn Guttman. SIGMOD 1984. pp 47-57.

10. [Beck90] The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. Norbert Beckmann, et al. SIGMOD 1990. pp 322-331.

11. [Mish92] Join Processing in Relational Databases. Priti Mishra, et al. ACM Computmg Surveys, Vol. 24, No 1, March 1992.

12. [Alkh02] Structural Joins: A Primitive for Efficient XML Query Pattern Matching. Shurug Al-Khalifa, et al. ICDE 02.

13. [Grae93] Query Evaluation Techniques for Large Databases. Goeta Graefe. ACM Computing Surveys,Vol. 25, No. 2, June 1993

14. [Seli79] Access Path Selection in a RDBMS. Pat Selinger. SIGMOD Conf., 1979, pp. 22-34.

---

[2] http://www.javaworld.com/javaworld/jw-04-1999/jw-04-xml.html

[3] http://www.rpbourret.com/xml/XMLAndDatabases.htm

15. [Jark84] Query Optimization in Database Systems. Matthias Jarke. Computing Surveys, Vol. 16, No. 2, June 1984.

Transaction Management and Concurrency Control
Week 6

16. [Gray76] Granularity of Locks and Degrees of Consistency in a Shared Data Base. J.N. Gray, et al. Modelling in Data Base Management Systems. North Holland Publishing Company, 1976.

17. [Agra87] Concurrency Control Performance Modeling: Alternatives and Implications. Rakesh Agrawal. et al. ACM Trans. on Database Systems, 12(4), 1987, pp. 609-654.

Database Research: Present and Future
Weeks 7, 8, 9 and 10

18. [Care93] The OO7 Benchmark. Michael J. Carey et al. SIGMOD Conf., 1993, pp. 12-21.

19. [Dewi] The Wisconsin Benchmark: Past, Present, and Future. DeWitt, et al.

20. [Schm01] The XML Benchmark Project. A.R. Schmidt, et al. Information Systems (INS) INS-R0103 April 30, 2001.

21. [Runa03] The Michigan Benchmark: Towards XML Query Performance Diagnostics. Kanda Runapongsa.

22. [Agra94] Fast Algorithms for Mining Association Rules. Agrawal, et al. VLDB Conf. 1994, pp. 487-499.

23. [Mohan07] A framework for XML Access Control – PhD Dissertation, Indiana University

24. [Fletcher06] Data Mapping as Search. George Fletcher and Catharine M. Wyss. EDBT 2006,

25. [Finance05] The case for access control on XML relationships. Beatrice Finance, Saia Medjdoub and Phileppe Pucheral, CIKM Conf 2005

26. [Yang 04] Secure XML publishing without information leakage in the presence of data inference Xiaochun Yang and Chen Li}, VLDB Conf 2004

27. [Silb96] Strategic Directions in Database Systems—Breaking Out of the Box. Avi Silberschatz. ACM Computing Surveys, Vol. 28, No. 4, December 1996

28. [Lamb91] The ObjectStore Database System. Charles Lamb, et al. CACM 34(10), 1991, pp. 51-63.

29. [Ston91] The POSTGRES Next Generation DBMS. M. Stonebraker, et al. CACM 34(10), 1991, pp. 78-92.