Final Exam Review Sheet
EE382V Data Engineering Fall ’10
Professor Daniel Miranker

Exam Date: Friday, 12/3/10, comprehensive 2 1/2 hour exam, closed book.

This review sheet is intended only as a study guide concerning the breadth of the exam. You are expected to know all the terminology presented as covered in class, the texts and the required supplemental reading. Again, individual terms and topics in this document are indicative of the breadth, not a comprehensive syllabus for the exam.

Reading:

In addition to the sections of the textbook corresponding to the lecture topics, it is understood that you will have read the required readings posted as posted on the web site and repeated here:

Text: Ch 7, 8.1 – 8.4, 9.1, 9.2, 13, 14.1, 14.2, 14.6, 14.7, 15
New since midterm: Ch. 16.1-16.6, 17.1-17.4, 18.1-18.4


Topics:

1. What is data modeling?
   a. Terminology
      i. Data model
      ii. Three schema model
         1. external schema
         2. conceptual schema
         3. internal schema
      iii. Entity, attribute, identifier, relation/association
      iv. Logical model, Physical Model, DDL
   b. What are the steps of a data modeling effort
      i. planning and analysis
      ii. conceptual design // logic without the details
      iii. logical design
      iv. physical design
      v. implementation
2. Basic Relational Database Concepts
   a. Schema(s)
   b. Content addressability
   c. Keys
      i. Candidate key
      ii. Primary key
      iii. Foreign key
      iv. Search/index key

3. Data Model Concepts
   a. Data modeling process
      i. Plan project
      ii. Determine requirements
      iii. Specify entities
   b. Entity, attribute, relation/association
   c. Weak and Strong Entities
   d. Relations/associations properties and implementation(s)
      i. Cardinality constraints
      ii. Aggregate
      iii. Inheritance (subtypes)
   e. Consistency Constraints & Triggers
      i. DBMS as an active manager of semantic correctness
      ii. Syntax

4. Disks and Data
   a. Physical properties of disk drives
   b. Two phase external sort

5. Indexing
   a. Methods
      i. B+-trees
      ii. R-trees, and other spatial partitioning methods
      iii. Bit-vector index methods
      iii. Metric space partitioning methods
   b. Secondary Indexes
      i. Applicability
      ii. Clustering

6. Query Systems
   a. Gross Structure
      i. Parsing
      ii. Logical Plan
      iii. Physical Plan
      iv. Optimizer
      v. Physical Operators
   b. Physical Operators
      i. Access Paths
         1. table_scan
         2. index_scan
      ii. Join Operators
1. Nested loops  
2. Merge join  
3. Hash-join  
4. Hybrid-hash join  

c. Estimated Query Cost  
   i. Estimating the cost of each operator  
   ii. Adorning a plan tree.  
   iii. Estimating the cost of a plan  

d. Optimization  
   i. Role of axioms/identified of the relational algebra  
   ii. Greedy rules e.g. pushing selects  
   iii. Dynamic programming method of optimizing join orders  
   iv. Use of a query graph  

7. Transactions  
   a. ACID properties  
      v. A…  
      vi. C…  
      vii. I…  
      viii. D…  
   b. Log-based recovery  
      i. Hardware organization  
         1. log  
         2. stable store  
      ii. Redo  
      iii. Undo  
      iv. Undo/redo  
      v. Role of commit  
   c. Concurrency Control  
      i. Schedule  
      ii. Serial Schedule  
      iii. Conflict/Bernstein Conditions  
      iv. Serializability  
      v. Precedence graph  
      vi. Serializability Theorem  
      vii. Locking  
         1. Two-phase locking  
         2. Escalading/upgrading locks  
            a. Shared locks  
            b. Granularity (basic ideas, not full implementation and correctness)  
      viii. Scheduler