Midterm 2 Review Sheet Spring '20
Database Systems

CS386D
Professor Daniel Miranker

Exam Date: April 23, Pending being able to configure the online test proctoring software, the exam will be closed book. You will have 1 hour and 15 minutes. You will be expected to start and complete the exam in the hour and 15 minutes between 4:00PM and 7:00PM Thursday, April 23.

In response to the Covid-19 situation, this is not a comprehensive exam, but it will overlap material from the first exam and in some cases, necessarily require knowledge of earlier material. Below is the syllabus for the exam.

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 reading. To be clear, individual terms and topics in this document are indicative of the breadth, *not* a comprehensive syllabus for the exam.

Reading:

Per the lecture schedule posted on Box.

Topics:

1. 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
    d. Basic Organization
        i) Data/index files
        ii) transaction log
        iii) SQL Engine
        iv) Role of RDBMS in a three-tier architecture

2. Constraints & Triggers
    a. Referential Integrity
    b. Attribute vs tuple level constraints
    c. Check constraints
    d. Triggers
        i. applications
        ii. integration with and implications of the transaction manager

3. Views
    a. Syntax/Semantics
    b. Use as a subroutine mechanism (nested queries)
    c. Use per logical restructing of a database, (re: external schema).
    d. Maintaining a materialized view
4. Disks and Data
    a. Physical properties of disk drives
    b. Two phase multiway merge sort
    c. ~~RAID~~

5. Indexing
    a. Methods (access paths)
        ~~i. B+ trees~~
        ~~ii. R-trees, and other spatial partitioning methods~~

~~iii. Bit-vector index methods~~
iii. Bloom filters

b. Secondary Indexes
   i. Applicability
   ii. Clustering
   iii. Effectiveness
      - Measures of effectiveness
      - Parameters, B(R), T(R), V(R, attr)

6. Query Systems
   a. Gross Structure
      - Parsing
      - Logical Plan
      - Physical Plan

   b. Physical Operators
      - Access Paths
         1. table_scan
         2. index_scan
      - Join Operators
         1. Nested loops
         2. Merge join
         3. Hash-join
         4. Hybrid-hash join
   c. Estimated Query Cost
      - Estimating the cost of each operator
      - Adorning a plan tree.
      - Estimating the cost of a plan, most notably
         - Impact of relational select and join operators
         - Measures wrt [text] I/O model and the number of rows
         - Architecture and organization (scope of the System Catalog)
   d. Optimization
      - Role of axioms/identified of the relational algebra
      - Greedy rules e.g. pushing selects
      - Dynamic programming method of optimizing join orders
      - Use of a query graph

9. Parallelism in Databases
   a. Machine organizations. e.g. shared-nothing
   b. Data Partitioning methods
      - round-robin
      - hashing
      - range based
   c. Application and use of hash partitioning in Cloud databases
      - Basically the ideas that appeared on the one midterm question
   d. data skew
   e. semi-join reduction
      - core idea
      - application to reduce communication in distributed/parallel query processing
      - approximate implementation using Bloom filters.
   f. Grace join