CSE 344
Midterm Review

July 26th

Midterm

- In class on Friday
- One sheet of notes, front and back
  - cost formulas also provided
- Practice exam on web site
- Good luck!

General Topics

- Databases
  - Motivations and definitions
- Relational Databases
  - SQL
  - Relational Algebra
  - Datalog
- Semi-structured Data
  - Motivations and definitions

General Topics

- Internals
  - Indexes
  - Physical plans/Cost Estimation
  - Disk I/o
- No Parallel DBs
  - (that will be on final exam)

Databases

- Motivations
  - Collections of related files
- Databases vs. DBMS
- What is stored?
- What is the DBMS' responsibility?

Databases

- Motivations
  - Collections of related files
- Databases vs. DBMS
- What is stored?
- What is the DBMS' responsibility?
  - Data storage and manipulation
  - Black box thought
  - Physical data independence

Relational Databases

- Motivations
  - Breaking away from singular flat files
  - Why/how do we break up data?
- Data model
  - Schemas and keys
  - Records and attributes
  - Attribute types/typing

Relational Databases

- Primary keys
  - What are the constraints?
  - When do we select keys?
  - Multiple keys
- Foreign keys
  - Constraints vs. Joining
- Keys across different data

SQL Structure

- Flat tables
  - First normal form
  - Breaking up data into multiple relations

SQL Code

- Create statements
  - Key declarations
  - Type declarations
  - Constraints: PK, FK, and general
- Insert/Delete statements
- Update statements
- Drop table

SQL Code

- Select
- From
- Where
- Group by
- Having
- Order by

SQL Code

- Distinct (and relation to group by)
- Inner vs. Outer Joining
  - Left/Right/Full
- Nested loop semantics
  - Cross product with selection
- Self joins
  - Produce companies that produce gadgets and cameras

SQL Code

- Aggregation
  - Count, sum, min, max, avg
- Null values
  - IS NOT null
  - Count(null)
- Where vs. Having

SQL Code

- Constructing Queries
  - FWGHOS (i.e., select is last)
- Subqueries
  - In Select (Single attribute projection)
  - In From (subquery AS, WITH AS)
  - In Where (EXISTS, IN, ANY)
  - Correlated vs. Non-correlated
  - Un-nesting
  - Finding the Witness

SQL Code

- Negation in subqueries
- Monotonicity
  - Definitions
  - Example
  - Difficulties and necessity of subqueries

Relational Algebra

- Set vs Bag semantics
  - Why bag?
- Query plans and RA expressions
- Operations (on relations, some with conditions)
  - Union, difference
  - Selection
  - Projection
  - Joins

Relational Algebra

- Operations (on relations, some with conditions)
  - Union, difference
  - Selection
  - Projection
  - Joins
  - Duplicate elimination
  - Grouping
  - Sorting

Relational Algebra

- Operations (on relations, some with conditions)
  - Union, difference
  - **Selection**
  - **Projection**
  - **Joins (remember your conditions)**
  - Duplicate elimination
  - **Grouping**
  - Sorting

Relational Algebra

- How do we know SQL and RA are equally expressive?
  - Translating one to the other
  - Multiple RA expressions possible for same query
  - DBMS optimization

Relational Algebra

- Producing RA expressions/trees
  - From queries
  - Visa-versa
- Bag vs. Set RA
  - Datalog is set semantic

Datalog

- Queries which cannot be defined in RA
    - Recursive queries
- Expressing RA expressions in datalog
    - Set semantics (procedural)
    - “Simple, concise, elegant”
- Fixed point semantics
    - Recursion builds from base case (empty)

Datalog

- Logical framework
- Explicitly defined intermediate results
- Terminology
  - Facts and Rules
  - Extensional vs. Intensional Predicates
  - Head and body
  - Head vs. Existential Variables
  - Unsafe rules

Datalog

- Writing Rules
  - Safety
  - Base cases
  - Aggregation and negation
  - Variable scope
  - Simple recursive queries
  - Converting from RA

Semistructured Data

- Motivations
  - Transactional vs. Analytical Workloads
  - Data distribution
  - Consistency
  - Partition vs. Replication
  - Key-value storage -> Document Storage

JSON

- Gives structure to data
- Objects and collections
- Self-describing
- Separate and less constrained than SQL++
- Nested structure (non-first normal form)

Asterix DB

- Document-based
- NoSQL
- Semi-structured
- Over JSON objects
  - Constraints (types, no duplicates)
- SQL++
  - Description vs. Manipulation

Asterix DB

- Dataverse
  - Database – set of data currently working with
- Types
  - UUID – auto generated
  - Null vs. Missing
  - Nested collections
  - Open v. Closed
  - Required v. Optional fields

Asterix DB

- Datasets
  - Relations
  - Defined over a type
  - Must have a key
- Indexes
  - Over particular attributes
  - Speeds up selections and joins

Asterix DB

- SQL++
  - Heterogeneity
  - Unnesting
  - Nesting/Aggregation and non-first normal
  - Multi-value join
    - data stores one to many instead of reverse
  - Can often be represented in SQL

Semistructured

- Distributed systems
- Short-term analysis
- Lower set-up costs
- Higher query costs (often)
- Higher query complexity
    - no free lunch... have to pay for costs of heterogeneity somewhere

Internals

- Physical Plans
  - Operators
    - Pipelining (selection, projection)
    - Joins
      - Nested Loop
      - Hash
      - Sorted merge
      - Index

Internals

- Physical Plans
  - Operators
    - Not discussed
      - Grouping/aggregation

Internals

- Physical Plans
  - Indexes
    - Clustered v. Unclustered
    - Hash v. B-Tree
    - When to apply
    - Benefit?

Internals

- Physical Plans
  - Cost estimation
    - Disk I/Os
    - Blocks and Tuples
    - Formulae (provided)
  - Tuple/block estimation
    - Selectivity factor

Questions

- That’s the material
- Things that will be on the exam
  - Relational data
    - schema design
    - queries in RA, SQL, Datalog
  - NoSQL
    - simplified data models
    - JSON and SQL++
  - Query optimization
    - cost estimation

Advice

- Look through the exam first
  - Try and do easiest questions first
  - Short answer questions are worth equal amounts, varying difficulty
  - Long exam, get easy points first
- Always be sure you understand the question

Advice

- Go through previous exams
  - Good judgement for questions
- Go through HW, WQ assignments
  - If I’ve asked you something before, I am certain that you should know how to do it
- Think about how null values/your assumptions impact the interpretation of the data