Introduction to Data Management CSE 344

Lecture 30: Final Review

Big Data

- Google's BigQuery announced yesterday Big Joins:
  - http://googleenterprise.blogspot.com/2013/03/bringing-simplicity-to-large-data.html
- Quote:
  - App Engine team reconciled app billing and usage information: Big JOIN of 2TB usage data with 10GB of configuration data in 60 seconds. (# servers omitted)
- Quote:
  - JOIN requires that the right-side table contains less than 8 MB of compressed data.
  - JOIN EACH allows join queries for tables of any size
- What are these joins?

The Final

- Wednesday, March 20th, 8:30-10:30
- In class
- Open notes and open books
- Review session: Saturday, 3/16, 10am, EEB 037

How To Study

- Go over the lecture notes
- Read the book
- Go over the homeworks
- Practice
  - Finals from past 344
  - Look at both midterms and finals from 444 past years: be careful because several questions do not apply to us!
  - Questions in the book
- Ask course staff questions or tomorrow in review session
- The goal of the final is to help you learn!

The Final

Entire class content is on the final!

But focus of questions on the final will be as follows:

1. SQL and Relational Query Languages (lectures 3-13)
2. XML (lectures 14-16)
3. Database design (lectures 17-20)
4. Views (lecture 21)
5. Transactions (lecture 22-24)
6. Parallel Databases (lecture 25-29)

3. SQL including Views

SQL

- SELECT-FROM-WHERE
- DISTINCT, ORDER BY, renaming of attributes
- INSERT, DELETE, UPDATE
- GROUP-BY and HAVING: *different* from WHERE (why ?)
- NULLs, outer joins
- Nested queries (subqueries)

Know the syntax

Know the semantics (nested loops !)

1. SQL and Relational Query Languages

SQL

- CREATE TABLE, plus constraints
- INSERT/DELETE/UPDATE

1. SQL and Relational Query Languages

- Relational algebra
- Relational calculus
- Nonrecursive datalog w/ negation

- Important translations:
  - Relational calculus → SQL
  - Relational calculus → Relational Algebra
  - SQL → Relational Algebra
  - Often convenient to first translate to datalog

2. XML

- Basic syntax: elements, attributes; well-formed v.s. valid document
- XPath
- XQuery

3. Database Design

E/R diagrams:

- Entities, attributes
- Relationships:
  - Many-many, many-one, one-one, exactly one
  - Multi-way relationships
- Inheritance, weak entity sets, union types
- Constraints in E/R diagrams
- Translation to relations

3. Database Design

Constraints in SQL

- Keys and Foreign Keys
- Attribute level constraints
  - Predicates on values
  - NOT NULL

3. Database Design

Conceptual Design

- Data anomalies
- Functional dependencies
  - Definition
  - Make sure you can check if a table satisfies a set of FDs
- Attribute closure
- Keys and Super keys
- Definition of BCNF
- Decomposition to BCNF

4. Views

- Types of views: virtual v.s. materialized views
- Definition and how to use them
- CREATE VIEW in SQL
- Query modification

5. Transactions

Transactions concepts

- Review ACID properties
- Definition of *serializability*
- Schedules, conflict-serializable and recoverable
- The four isolation levels in SQL
- Concurrency control using locks
  - SQLite and SQLServer examples
- Phantoms, dirty reads, and other problems
- Deadlocks
- Transactions in SQL

6. Parallel Data Processing

Parallel databases:

- Speedup/scaleup
- Shared memory, shared disk, shared nothing
- Horizontal data partition: block, hash, range
- How to implement simple algorithms: group-by, join
- How to execute a complete query in parallel

6. Parallel Data Processing

MapReduce

- Functions: map, (combine,) reduce
- Terminology: chunk, map job / reduce job; map task / reduce task; server (instance); failed server
- Basic implementation of MR
- Dealing with server failures and stragglers
- How to express simple computations in MapReduce

You will not be asked to write a Pig Latin query, but should have some basic understanding of how queries are implemented over MapReduce