Introduction to Data Management
CSE 344

Lecture 26: Data Integration

Final Exam

- Wednesday, Dec 14, 8:30am-10:20am, in class
- Content: Everything
  - But no questions on the relational calculus
- Open books and open notes
  - You can bring any notes that you want including slides, assignments, old exams, stuff from the web, ...
  - BUT NO laptops, phones, or other devices

Possible Questions about Last Few Lectures

- DBMS-as-a-service: You can expect some sort of qualitative question that will ask you to
  - Discuss the benefits and challenges of offering DBMSs as a service in the cloud.
  - Discuss some similarities and differences between relational and NoSQL systems or the motivation for such systems.
  - Something else along those lines.
  - NO specific questions about the various clouds
- Data integration and data cleaning
  - You can similarly expect a qualitative question on these topics

How To Study

- Go over the lecture notes
- Read the book
- Go over the homeworks
- Practice
  - Practice webquiz will be posted Wednesday night
  - Look at the final from 344 Spring 2011
  - Look at both midterms and finals from 444 past years: be careful because several questions do not apply to us!
- Ask Kristi and me questions
- The goal of the final is to help you learn!

Today: Data Integration

- Goal:
  - Data is available in multiple distinct databases
  - Want to ask queries across all these databases
- When is data integration needed?
  - Two companies merge
  - Want to get legacy databases to talk to each other
  - Want to analyze data produced by different sources
  - Want to combine data from different websites

Data Integration Challenges (1/2)

- Each database could be in a different type of DBMS (different data model, query language, etc.)
  - Relational, semi-structured, NoSQL
- Schema heterogeneity
  - S1: Employer(ID, name, address, position, salary)
  - S2: Worker(EID, name, address) Position(PID,salary,from,until)
- Data type heterogeneity
  - Employee ID could be a string or an integer
- Value heterogeneity
  - The “cashier” position could be called “cashier” or “associate”
Data Integration Challenges (2/2)

- **Semantic heterogeneity**
  - Most difficult to manage
  - E.g., salary is hourly salary before tax
  - Or salary is net, weekly salary with lunch allowance
- **Data integration is a very, very, very difficult pb!**

Data Integration Approaches

- **Federated databases**
  - Each source remains independently administered
  - One source can call on others to supply info
- **Centralized warehouse**
  - Data from source is extracted-transformed-loaded into a single, centralized database
  - Data is refreshed periodically (so data is not 100% up-to-date)
- **Mediator**
  - Virtual database on top of others
  - Takes query as input and rewrites it in terms of queries over the other databases, then synthesizes the answer

Creating a Data Warehouse

- **Extract** data from distributed operational databases
  - Can do this by running a query over the data source
- **Clean** to minimize errors and fill in missing information
  - We will discuss cleaning next lecture
- **Transform** to reconcile semantic (and other) mismatches
  - Performed by defining views over the data sources
- **Load** to materialize the above defined views
  - Build indexes and additional materialized views
- **Refresh** to propagate updates to warehouse periodically
  - Update the warehouse incrementally

Executing Queries over Mediator

- **Wrappers more complex than with warehouses**
  - Need to execute all sorts of queries, not just extract
  - One approach is to define templates
  - A template is a parameterized query
    - `select * from EmployeeMed where position='$p'`
- **Query optimization at the mediator is a challenge**
  - Wrapped data sources can be seen as views
  - How can I answer the given query using these views?
- **Data sources can exhibit bad and variable performance**
  - May want a more dynamic query plan: process data as it arrives

Dirty Data

- **Another challenge with data integration**
  - Often hard to decide if two records represent the same entity or not
  - E.g., John Doe from 1234 56th ave NE, Seattle
  - Vs. J. R. Doe from 1234 56th ave NE, Seattle
  - Vs John Doe from 789 108th St., Bellevue
- **Even without data integration, data often dirty**
  - Missing values, duplicates, odd characters, etc.