Introduction to Database Systems
CSE 414

Lecture 29: Data Integration

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: Employee(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 problem!

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
- 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*

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.

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 to 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

Not the end of the story…

… but all we have time for in 10 weeks

(But we managed to cover a lot)

Review / wrapup / evaluations on Friday

See you then!