Data Integration

Alon Halevy
Google Inc.
University of Aalborg
September, 2007

Introduction

What is Data Integration and Why is it Hard?

DBMS: it’s all about abstraction

Logical vs. Physical; What vs. How.

Students: T ID: Ssn Name Category
123-45-6789 Charles undergrad
234-56-7890 Dan grad

Courses: C ID: Name Quarter
CSE444 Databases fall
CSE541 Operating systems winter

Data Integration: A Higher-level Abstraction

Query

Mediated Schema

Independence of:
• source & location
• data model, syntax
• semantic variations

Semantic Mappings

SELECT C.name
FROM Students S, Takes T, Courses C
WHERE S.name = “Mary” and S.ssn = T.ssn and T.cid = C.cid

Application Area 1: Business

EII Apps:
CRM
ERP
Portals

Application Area 2: Science

Hundreds of biomedical data sources available; growing rapidly!

50% of all IT $$$ spent here!
Application Area 3: The Web

FullServe Corporation

Employees
- FullTimeEmp
- Hire
- TempEmployees

Training
- Courses
- Enrollments

Sales
- Products
- Sales

Resumes
- Interview
- CV

Services
- Services
- Customers
- Contracts

HelpLine
- Calls

EuroCard Corporation

Employees
- Employees
- Hire

Credit Cards
- Customer
- CustDetail

Resumes
- Interview

HelpLine
- Calls

Heterogeneity 101

Employees
- FullTimeEmp
  - ssn, empId, firstName, middleName, lastName
- Hire
  - empId, hireDate, recruiter
- TempEmployees
  - ssn, hireStart, hireEnd

Employees
- ID, firstNameMiddleInitial, lastName
- Hire
  - ID, hireDate, recruiter

Find all employees (making over $100K)

Customer Call Center

Customer: I lost my credit card!
Agent: would you like to buy an espresso machine?

Sales
- Products
- Sales

Credit Cards
- Customer
- CustDetail

Challenges & Opportunities

- Create a (useful) web site for tracking services
- Collaborate with third parties
  - E.g., create branded services
- Comply with government regulations
  - Find “risky” employees
- Business intelligence
  - What’s really wrong with our products?
The Deep Web

- Millions of (good) forms out there
- Each form has its own special interface
  - Hard to explore data across sites.
- Goal (for some domains):
  - A single interface into a multitude of deep-web sources.

The Semantic Web

- Knowledge sharing at web scale.
- Web resources are described by ontologies:
  - Rich domain models; allow reasoning.
  - RDF/OWL are the emerging standards
  - OWL-lite may actually be useful.
- Issues:
  - Too complex for users?
  - Killer apps?
  - Scalability of reasoning?
- Proposal: let's build the SW bottom up.

Goal of Data Integration

- Uniform query access to a set of data sources
- Handle:
  - Scale of sources: from tens to millions
  - Heterogeneity
  - Autonomy
  - Semi-structure
Why is it Hard?

- **Systems-level reasons:**
  - Managing different platforms
  - SQL across multiple systems is not so simple
  - Distributed query processing

- **Logical reasons:**
  - Schema (and data) heterogeneity

- **‘Social’ reasons:**
  - Locating and capturing relevant data in the enterprise;
  - Convincing people to share (data fiefdoms)
  - Security, privacy and performance implications.

Setting Expectations

- Data integration is *AI-Complete*.
  - Completely automated solutions unlikely.

Goal 1:

- Reduce the effort needed to set up an integration application.

Goal 2:

- Enable the system to perform gracefully with uncertainty (e.g., see the web)

Relational Terminology

- **Relational schemas**
  - Tables, attributes

- **Relation instances**
  - Sets (or multi-sets) of tuples

- **Integrity constraints**
  - Keys, foreign keys, inclusion dependencies

SQL (very basic)

**Interview:**

- candidate, date, recruiter, hireDecision, grade

**EmployeePerf:**

- empID, name, reviewQuarter, grade, reviewer

```sql
select recruiter, candidate
from Interview, EmployeePerf
where recruiter=name AND grade < 2.5
```
SQL (w/ aggregation)

EmployeePerf:
  empID, name, reviewQuarter, grade, reviewer

\[
\begin{align*}
\text{select} & \quad \text{reviewer}, \text{Avg}\text{\,(grade)} \\
\text{from} & \quad \text{EmployeePerf} \\
\text{where} & \quad \text{reviewQuarter} = "1/2007"
\end{align*}
\]

Conjunctive Queries

\[
\begin{align*}
Q(R, C) & :- \\
& \text{Interview}(X, D, Y, H, F), \text{EmployeePerf}(E, Y, T, W, Z), \\
& W < 2.5.
\end{align*}
\]

\[
\begin{align*}
\text{select} & \quad \text{recruiter}, \text{candidate} \\
\text{from} & \quad \text{Interview, EmployeePerf} \\
\text{where} & \quad \text{recruiter} = \text{name} \ \text{AND} \ \text{grade} < 2.5
\end{align*}
\]

Unions of Conjunctive Queries

\[
\begin{align*}
Q(R, C) & :- \\
& \text{Interview}(X, D, Y, H, F), \text{EmployeePerf}(E, Y, T, W, Z), \\
& W < 2.5.
\end{align*}
\]

\[
\begin{align*}
Q(R, C) & :- \\
& \text{Interview}(X, D, Y, H, F), \text{EmployeePerf}(E, Y, T, W, Z), \\
& \text{Manager}(y), W > 3.9.
\end{align*}
\]

Datalog (recursion)

\[
\begin{align*}
\text{Database: edge}(X, Y) \\
\text{Path}(X, Y) & :- \text{edge}(X, Y) \\
\text{Path}(X, Y) & :- \text{edge}(X, Z), \text{path}(Z, Y)
\end{align*}
\]

Warmup Exercise

Virtual Data Integration Architecture
Mediated Schema

Source Descriptions

Wrappers

Mediation Languages

Query Processing

Woody Allen Comedies in NY
Outline (1)

- Mediation languages:
  - theoretical foundations (containment, answering queries using views)
  - Creating schema mappings
  - Query processing
    - Adaptive query processing
  - XML and its role in data integration

Outline (2)

- Other architectures for data integration
  - Warehousing, data exchange, p2p dbms
- Advanced (i.e., current) topics:
  - Dataspaces
  - The deep web
  - Uncertainty in data integration