CIS 602: Provenance & Scientific Data Management

Database Provenance

Dr. David Koop
Relational Databases Review

- Databases composed of tables
  - Each table has rows and columns
  - Key is a unique identifier for the row
  - Foreign keys help link one table to another

- Database schemas define the structure of the database
  - Usually defined before the database is used
  - CREATE TABLE Employee(empid char(12), dept varchar(31));

Source database $D$:

<table>
<thead>
<tr>
<th>empid</th>
<th>dept</th>
</tr>
</thead>
<tbody>
<tr>
<td>e977</td>
<td>CS</td>
</tr>
<tr>
<td>e132</td>
<td>EE</td>
</tr>
<tr>
<td>e657</td>
<td>BME</td>
</tr>
</tbody>
</table>

Department

<table>
<thead>
<tr>
<th>deptid</th>
<th>budget</th>
</tr>
</thead>
<tbody>
<tr>
<td>BME</td>
<td>1200K</td>
</tr>
<tr>
<td>CS</td>
<td>670K</td>
</tr>
<tr>
<td>EE</td>
<td>890K</td>
</tr>
<tr>
<td>MATH</td>
<td>230K</td>
</tr>
</tbody>
</table>

Output of $Q(D)$:

<table>
<thead>
<tr>
<th>empid</th>
<th>dept</th>
<th>budget</th>
</tr>
</thead>
<tbody>
<tr>
<td>e657</td>
<td>BME</td>
<td>1200K</td>
</tr>
<tr>
<td>e977</td>
<td>CS</td>
<td>670K</td>
</tr>
<tr>
<td>e132</td>
<td>EE</td>
<td>890K</td>
</tr>
</tbody>
</table>
Relational Databases Review

- Most relational database queries are made using SQL (Structured Query Language)
- Based on relational algebra
  - Select, Project, Join
  - SELECT
  - FROM
  - WHERE
  - GROUP BY
  - HAVING
  - ORDER BY

Query Q:

```sql
SELECT e.empid, e.dept, d.budget
FROM Employee e, Department d
WHERE e.dept = d.deptid
```
Reading Response Reminders

• Sent emails to entire class about plagiarism
  - *Your* summaries are meant to be *your* own views on the paper, to show *your* understanding of the material

• Sent individual emails with reviews of the first four responses

• Generally:
  - Work on writing, linking sentences, elucidating points
  - Work on focusing on key ideas and potential issues in critiques
  - Check spelling, grammar, punctuation!
Course Project Proposal

- Due October 2 at 12pm, also a few slides for October 2 class

- Requirements:
  - Standard metadata: Title, Authors, Date
  - Introduction with description of the project goals
  - Motivation: why did you choose this project and how does it tie into the topics of the class
  - Background: what do we need to understand as background
  - Design: what are the components and for your project
  - Implementation Details: (language, platform(s), system requirements)
  - Project plan: what are the different components and what is your timeline for implementing the different pieces?
Reading Quiz
Provenance in Databases: Past, Current, and Future

W.-C. Tan

Presented by: Zachary Sylvia
Towards Integrating Workflow and Database Provenance

F. Chirigati and J. Freire
IPAW 2012
Reminders

- Course Project Proposals
  - If you would like me to look over a draft of proposal, please email it to me
  - Due October 2 at 12pm, small presentations during class