Hippocratic Databases

Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu
IBM Almaden Research Center, VLDB 2002

George W. Boulos
gwf5@pitt.edu
October 14 2009
Motivational Example

- Online Book Seller (Mississippi):
  - Purchase (shipping address, name, email, credit card)
  - Book Recommendations (books purchased, email)
  - Purchase Circles (location)

- Alice:
  - Purchase

- Bob:
  - Purchase
  - Recommendations
“Privacy is the right of individuals to determine for themselves when, how and to what extent information about them is communicated to others.”

Alan Westin, Professor Emeritus of Public Law and Government, Columbia University.

We need a new Database architecture to support the new needs for preserving individual “Privacy”.

We will cover the next topics:
  • Current Database Systems
  • Principles and Regulations for preserving privacy
  • Suggested Design
  • Challenges Discussion
  • Q&A
Current Databases Properties

1. **Manage persistent data.**
2. **Access a large amount of data efficiently.**
3. **Data model** support (at least one).
4. **High-level languages** that allow the user to define the structure of data, access data, and manipulate data.
5. **Transaction management:** the capability to provide correct, concurrent access to the database by many users at once.
6. **Access control:** the ability to deny access to data by unauthorized users and the ability to check the validity of the data.
7. **Resiliency:** the ability to recover from system failures without losing data.
Statistical Databases

• Query Restriction:
  • restricting the size of query results.
  • keeping audit trails of all answered queries and constantly checking for possible compromises.
  • suppression of data cells of small size, and clustering entities into mutually exclusive atomic populations.

• Data Perturbation:
  replacing the original database by a sample from the same distribution, adding noise to the values in the database.
Secure Databases

• Data transmitted and stored securely
• Supports Security Levels:
  – A query with a lower level of security cannot read a data item requiring a higher level of clearance.
  – A higher security query cannot write a lower security data item.

*Does this scheme remind you of anything?*
Principles of Hippocratic Databases

1. **Purpose Specification:** For personal information stored in the database, the purposes for which the information has been collected shall be associated with that information.

2. **Consent:** The purposes associated with personal information shall have consent of the donor of the personal information.

3. **Limited Collection:** The personal information collected shall be limited to the minimum necessary for accomplishing the specified purposes.
4. **Limited Use:** The database shall run only those queries that are consistent with the purposes for which the information has been collected.

5. **Limited Disclosure:** The personal information stored in the database shall not be communicated outside the database for purposes other than those for which there is consent from the donor of the information.

6. **Limited Retention:** Personal information shall be retained only as long as necessary for the fulfillment of the purposes for which it has been collected.

7. **Accuracy:** Personal information stored in the database shall be accurate and up-to-date.
8. **Safety**: Personal information shall be protected by security safeguards against theft and other misappropriations.

9. **Openness**: A donor shall be able to access all information about the donor stored in the database.

10. **Compliance**: A donor shall be able to verify compliance with the above principles. Similarly, the database shall be able to address a challenge concerning compliance.
Hippocratic Databases Design
### Figure 2: Privacy Metadata Schema

<table>
<thead>
<tr>
<th>table</th>
<th>attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>privacy-policies</td>
<td>purpose, table, attribute, {external-recipients}, retention</td>
</tr>
<tr>
<td>privacy-authorizations</td>
<td>purpose, table, attribute, {authorized-users}</td>
</tr>
</tbody>
</table>

### Figure 3: Database Schema

<table>
<thead>
<tr>
<th>table</th>
<th>attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>customer</td>
<td>purpose, customer-id, name, shipping-address, email, credit-card-info</td>
</tr>
<tr>
<td>order</td>
<td>purpose, customer-id, transaction-id, book-info, status</td>
</tr>
</tbody>
</table>

### Figure 4: Privacy-Policies Table

<table>
<thead>
<tr>
<th>purpose</th>
<th>table</th>
<th>attribute</th>
<th>external-recipients</th>
<th>retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>purchase</td>
<td>customer</td>
<td>name</td>
<td>{delivery-company, credit-card-company}</td>
<td>1 month</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>shipping-address</td>
<td>{delivery-company}</td>
<td>1 month</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>email</td>
<td>empty</td>
<td>1 month</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>credit-card-info</td>
<td>{credit-card-company}</td>
<td>1 month</td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>book-info</td>
<td>empty</td>
<td>1 month</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>name</td>
<td>empty</td>
<td>3 years</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>shipping-address</td>
<td>empty</td>
<td>3 years</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>email</td>
<td>empty</td>
<td>3 years</td>
</tr>
<tr>
<td>recommendations</td>
<td>order</td>
<td>book-info</td>
<td>empty</td>
<td>10 years</td>
</tr>
<tr>
<td>purchase-circles</td>
<td>customer</td>
<td>shipping-address</td>
<td>empty</td>
<td>1 year</td>
</tr>
<tr>
<td>purchase-circles</td>
<td>order</td>
<td>book-info</td>
<td>{aggregated-all}</td>
<td>1 year</td>
</tr>
<tr>
<td>purpose</td>
<td>table</td>
<td>attribute</td>
<td>authorized-users</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------</td>
<td>------------------</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>customer-id</td>
<td>all</td>
<td></td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>name</td>
<td>{ shipping, charge, customer-service }</td>
<td></td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>shipping-address</td>
<td>{ shipping }</td>
<td></td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>email</td>
<td>{ shipping, customer-service }</td>
<td></td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>credit-card-info</td>
<td>{ charge }</td>
<td></td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>customer-id</td>
<td>all</td>
<td></td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>transaction-id</td>
<td>all</td>
<td></td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>book-info</td>
<td>{ shipping }</td>
<td></td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>status</td>
<td>{ shipping, customer-service }</td>
<td></td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>customer-id</td>
<td>all</td>
<td></td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>name</td>
<td>{ registration, customer-service }</td>
<td></td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>shipping-address</td>
<td>{ registration }</td>
<td></td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>email</td>
<td>{ registration, customer-service }</td>
<td></td>
</tr>
<tr>
<td>recommendations</td>
<td>order</td>
<td>customer-id</td>
<td>{ mining }</td>
<td></td>
</tr>
<tr>
<td>recommendations</td>
<td>order</td>
<td>transaction-id</td>
<td>{ mining }</td>
<td></td>
</tr>
<tr>
<td>recommendations</td>
<td>order</td>
<td>book-info</td>
<td>{ mining }</td>
<td></td>
</tr>
<tr>
<td>purchase-circles</td>
<td>customer</td>
<td>customer-id</td>
<td>{ olap }</td>
<td></td>
</tr>
<tr>
<td>purchase-circles</td>
<td>customer</td>
<td>shipping-address</td>
<td>{ olap }</td>
<td></td>
</tr>
<tr>
<td>purchase-circles</td>
<td>order</td>
<td>customer-id</td>
<td>{ olap }</td>
<td></td>
</tr>
<tr>
<td>purchase-circles</td>
<td>order</td>
<td>book-info</td>
<td>{ olap }</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Privacy-Authorizations Table
Hippocratic Database Design Cont.

• Hippocratic Databases stores the “Purpose” for which the information has been collected, which is analogous to Role.

• Privacy Metadata defines:
  • External Recipients
  • Retention Period
  • Authorized Users

We can see that the traditional Logic Based Engine in Access control lists was replaced by metadata tables, Do you think that is better or worst?
Data Collection

• Matching Privacy Policy with User Preferences
  • Before the user provides any information, the Privacy Constraint Validator checks whether the business’ privacy policy is acceptable to the user.

• Data Insertion
  • Saving data securely.

• Data Processing
  • The Data Accuracy Analyzer may run some data cleansing functions against the data to check for accuracy.
Queries

• Before Query
  • Execution A query is only allowed to run if the set of authorized users for that purpose in the privacyauthorization table includes the user who issued the query.

• During Query
  • the Record Access Control ensures that only records whose purpose attribute includes the query’s purpose will be visible to the query.

• After Query
  • The Query Intrusion Detector is run on the query results to spot queries whose access pattern is different from the usual access pattern for queries with that purpose and by that user.
Retention

The *Data Retention Manager* deletes *data items* *that have* outlived their purpose. If a certain data item was collected for a set of purposes, it is kept for the retention period of the purpose with the highest retention time.
Data Collection Analyzer

• examines the set of queries for each purpose to determine if any information is being collected but not used.

• It also determines if data is being kept for longer than necessary, and whether people have unused authorizations to issue queries with a given purpose.

“Efficiency - though it will continue to be important- may not be the central focus of Hippocratic Databases” section 2

Do you think that there could be a commercially available database that was not designed optimally for performance?!
Platform for Privacy Preferences (P3P)

• P3P provides a way for a Web site to encode its data-collection practices in a machine-readable XML format known as a P3P policy.

• we can take P3P policies, process them through the privacy metadata processor, and generate the corresponding data structures.
Challenges

• Language for Privacy Policies and Preferences
• Efficiency
• Limited Collection
• Limited Disclosure
• Limited Retention
• Safety
• Openness
• Compliance
  – Universal Logging
  – Tracking Privacy Breaches
Strength and Weaknesses

→Strengths:
   →Proper Design.
   →Paper fairly achieves it is goal of defining Hippocratic Database.

→Weaknesses:
   →Yet there exists some redundancy in the authorization model described.
   →The paper does not cover several basic scenarios a database undergoes, like editing the model.
Discussion Point

→ Hippocratic *Database define* some privacy constraints and if the user agrees on them it proceeds.

→ I think the *user should define* his privacy preference and the database should handle it accordingly.
Thank you