Basics

- Data warehouse is an integrated repository derived from multiple distributed source databases.
- Created by replicating or transforming source data to new representation.
- Some data can be web-database or regular databases (relational, files, etc.).
- Warehouse creation involves reading, cleaning, aggregating, and storing data.
- Warehouse data is used for strategic analysis, decision making, market research types of applications.
- Open access to third party users.

Examples:

- Human genome databases.
- Drug-drug interactions database created by thousands of doctors in hundreds of hospitals.
- Stock prices, analyst research.
- Teaching material (slides, exercises, exams, examples).
- Census data or similar statistics collected by government.

Ideas for Security

- Replication
- Aggregation and Generalization
- Exaggeration and Mutilation
- Anonymity
- User Profiles, Access Permissions

Anonymity

One can divulge information to a third party without revealing where it came from and without necessarily revealing the system has done so.

- User privacy and warehouse data privacy.
- User does not know the source of data.
- Warehouse system does not store the results and even the access path for the query.
- Separation of storage system and audit query system*.
- Non-intrusive auditing and monitoring.
- Distribution of query processing, logs, auditing activity.
- Secure multi-party computation.
- Mental poker (card distribution).

* Research project of Atallah and Prabhakar at Purdue.

- Witness (Permission Inference)

  User can execute query Q if there is an equivalent query Q′ for which the user has permission. Security is on result and not computation.
- Create views over mutually suspicious organizations by filtering out sensitive data.

Similarity Depends on Application

- Two objects might be similar to a K-12 student, but not a scientist.
- 1999 and 1995 annual reports of the CS department might be similar to a graduate school applicant, but not to a faculty applicant.

Goal: Use ideas of replication to provide security by using a variety of similarity criterion

Goal: Different QoS to match different classes of users.

Similarity Based Replication*

SOME DEFINITIONS:

- ***Distinct functions*** used to determine how similar two objects are (Distinct Preserving Transformations).
- ***Precision***: fraction of retrieved data as needed (relevant) for the user query.
- ***False Positive***: object retrieved that is similar to the data needed by query, but it is not.
- ***False Negative***: object is needed by the query, but not retrieved.

* Bhargava/Annamalia, Defining Data Equivalence, IDPT, 1996

Access Permission*

- Information permission (system-wide)
  - (employee salary is releasable to payroll clerks and cost analyst).
- Physical permission (local)
  - (cost analysts are allowed to run queries on the warehouse).

* Rosenthal & Sciore, DMDW 2000 (view security…) SOL extensions.

Cooperation Instead of Autonomy in Warehouse*

- In UK, the Audit Commission estimated losses of the order of $2 billion.
- Japanese Yakuza made a profit of $7 billion.
- A secure organization needs to secure data, as well as it's interpretation.

  (Integrity of data OK, but the benefit rules were interpreted wrong and misapplied.)

  $\Rightarrow$ Interpretation Integrity

* Dhillon & Backhouse, Inf. Syst. Mgt., CACM – July 2000.

Extensions to the SQL Grant/Revoke Security Model*

- Limitation is a generalization of revoke.
- Limitation Predicates should apply to only paths (reduces chance of inadvertent & malicious denial of service).
- One can add either limitation or reactivation, or both.
- Limitation respects lines of authority.
- Flexibility can be provided to limitation.

---

* Rosenthal & Sciore, IFIP Conf. On Security, 2000.

- Cascade Revoke, Reactivation Without Cascase, Bertino/Jajodia/Samarati, ACM TIS, 99.

Aggregation and Generalization

- Summaries, Statistics
  - (over large or small set of records)
  - (various levels of granularity)
- Graphical image with numerical data.
- Reduce the resolution of images.
- Approximate answers
  - (real-time vs. delayed quotes, blood analysis results)
- Inherit access to related data.

Dynamic

- Authenticate users dynamically and provides access privileges.
  - Mobile agent interacts with the user and provides authentication and personalized views based on analysis and verification.
- Rule-based interaction session.
- Analysis of the user input.
- Determination of the user's validity and creating a session id for the user and assignment of access permission.

Exaggeration and Misleading

- Give low or high range of normal values. Initially (semantically normal).
- Partially incorrect or difficult to verify data. Quality improves if security is assured.
- Give old data, check damage done, give better data.
- Projected values than actual values.

User Profile

- User profiles are used for providing different levels of security.
- Each user can have a profile stored at the web server or at third party server.
- User can change profile attributes at run-time.
- User behavior is taken into account based on past record.
- Mobile agent accesses the web page on behalf of the user and tries to negotiate with web server for the security level.

User Profile

- Personal category
  - personal identifications; name, dob, ss, etc.
- Data category
  - document content; keywords
  - document structure; audio/video, links
  - source of data
- Delivery data – web views, e-mail
- Secure data category

Static

- Predefined set of user names, domain names, and access restrictions for each
  - (restricted & inflexible)
- Virtual view, Materialized view, Query driven
- Build user profiles and represent them
  - past behavior
  - feedback
  - earlier queries
  - type, content and duration