Data Disclosure Technology for the Grid

Tyrone W A Grandison PhD

Data Disclosure

- **The release of information for processing**
- **Security & privacy concerns critical for Grid**
- **One way to address this:**
  - Hippocratic Databases

GOAL

Create a new generation of information systems that protect the privacy, security, and ownership of data while not impeding the information flow.

Disclosure Control

Database-enforced disclosure control at cell-level of an organization's data policy and user preferences

Privacy Preserving Data Analytics

Preserve privacy at the individual level, while still building accurate data mining models at the aggregate level

Sovereign Information Sharing

Selective, minimal sharing across database separated due to statutory, competitive or security reasons

Compliance Auditing

Audit if data has been disclosed in violation of the specified policy to help enterprises cope with increasing pace of legislations

What is Hippocratic Database technology?

Technology that facilitates automated non-intrusive, high-performance, fine-grained data disclosure (at the database level).

Functional Components:

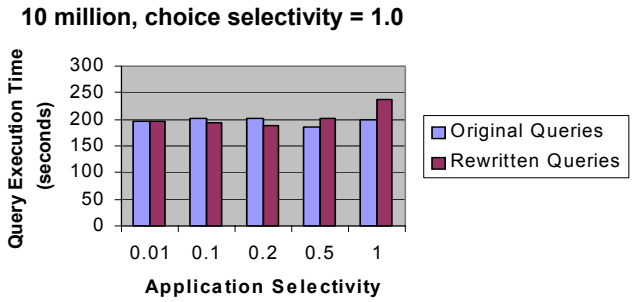
- Active Enforcement ***– enabling the database to reveal only data compliant with policy.***
- Eunomia Compliance Auditing ***– enabling verification and monitoring of compliance with policy (e.g., legislation, privacy, security).***
- Sovereign Information Integration ***– enabling two parties to securely and privately share common information without a third party involved.***
- Order Preserving Encryption ***– enabling the protection of data from medium theft, while allowing encrypted data to be usable.***
- Watermarking Databases ***– deterring data theft and asserting ownership of pirated copies.***
- Privacy Preserving Data Mining ***– preserving privacy at the individual level, while enabling the construction of accurate data mining models at the aggregate level.***
- BA k-anonymity ***– enable data release, while preventing linking attacks and preserving data integrity***

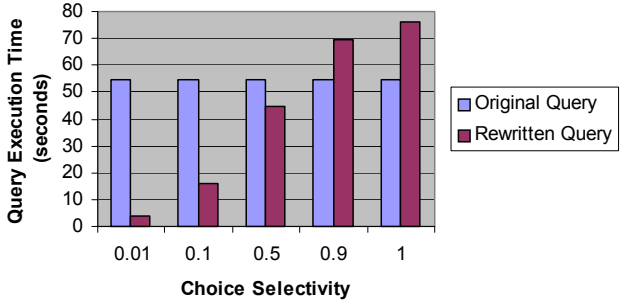
Active Enforcement

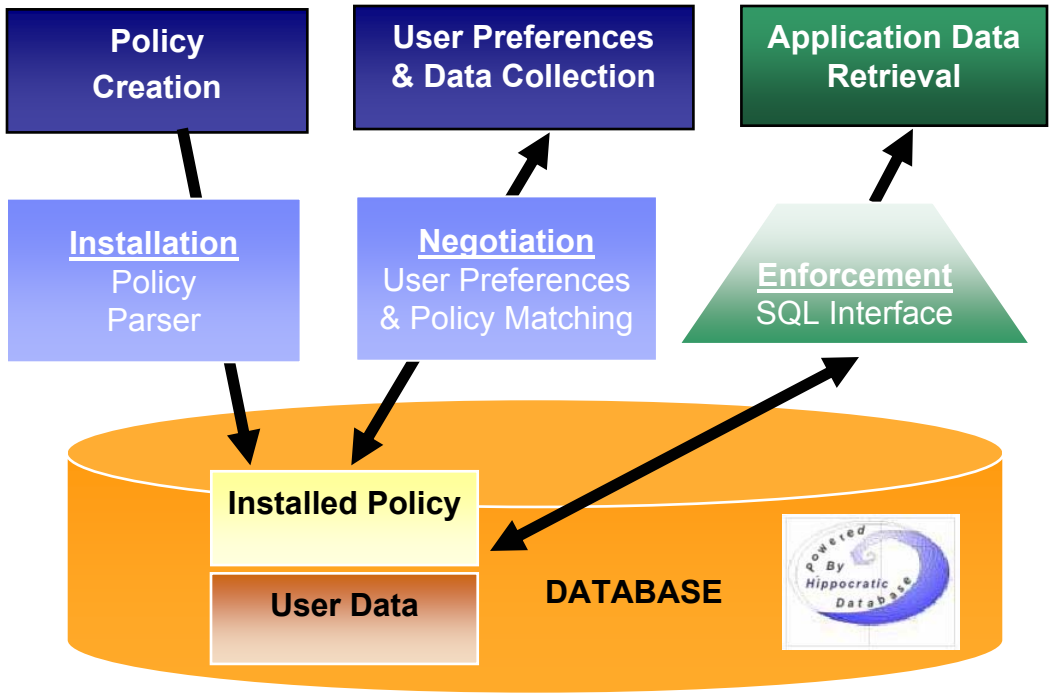
- Database-enforced disclosure control at cell-level of an organization's data policy and user preferences.
- Applications do not require any modification.
- Database agnostic, does not require any change in the database engine.

- Implementation intercepts and rewrites incoming queries to factor in policy, user choices, and context (e.g. purpose).
- Rewritten queries benefit from all the optimizations and performance enhancements provided by underlying engine (e.g. parallelism).
- Demo available for HIPAA enforcement

#	Name	Age	Phone
1	Adams	25	111-1111
3	-	-	333-3333
4	Daniels	40	-



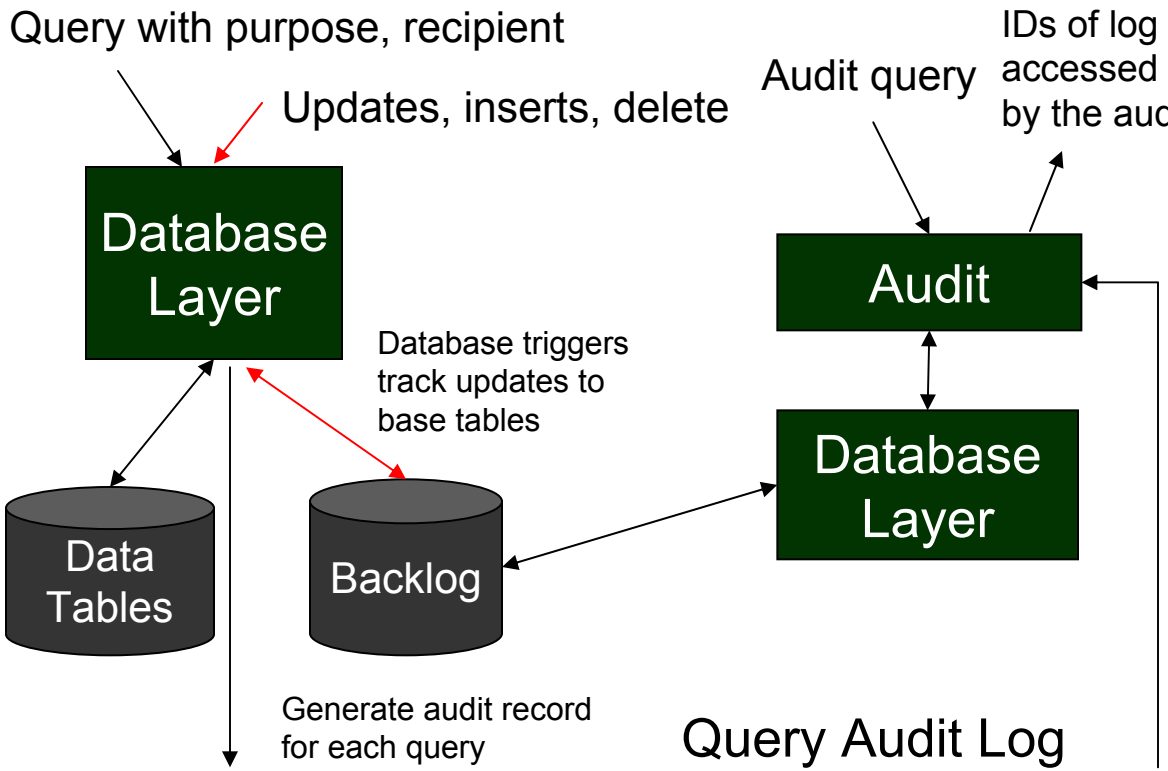




Compliance Auditing

- **Problem: Many laws require companies to account for their handling of sensitive or secured information.**
- **How does a company ensure and show that it has complied with such requirements?**
  - Existing database systems and tools provide only rudimentary query logging which is rarely sufficient.
  - Other add-on applications can also log query results, but this has a huge performance impact and still does not reveal certain disclosures of sensitive information.
- **Solution: An efficient auditing system that tracks disclosures down to the cell level in the database.**
  - Allow determining precisely who accessed designated data, for what purpose, when it was accessed, and what changes were made.
  - With minimal impact on the company's operations.

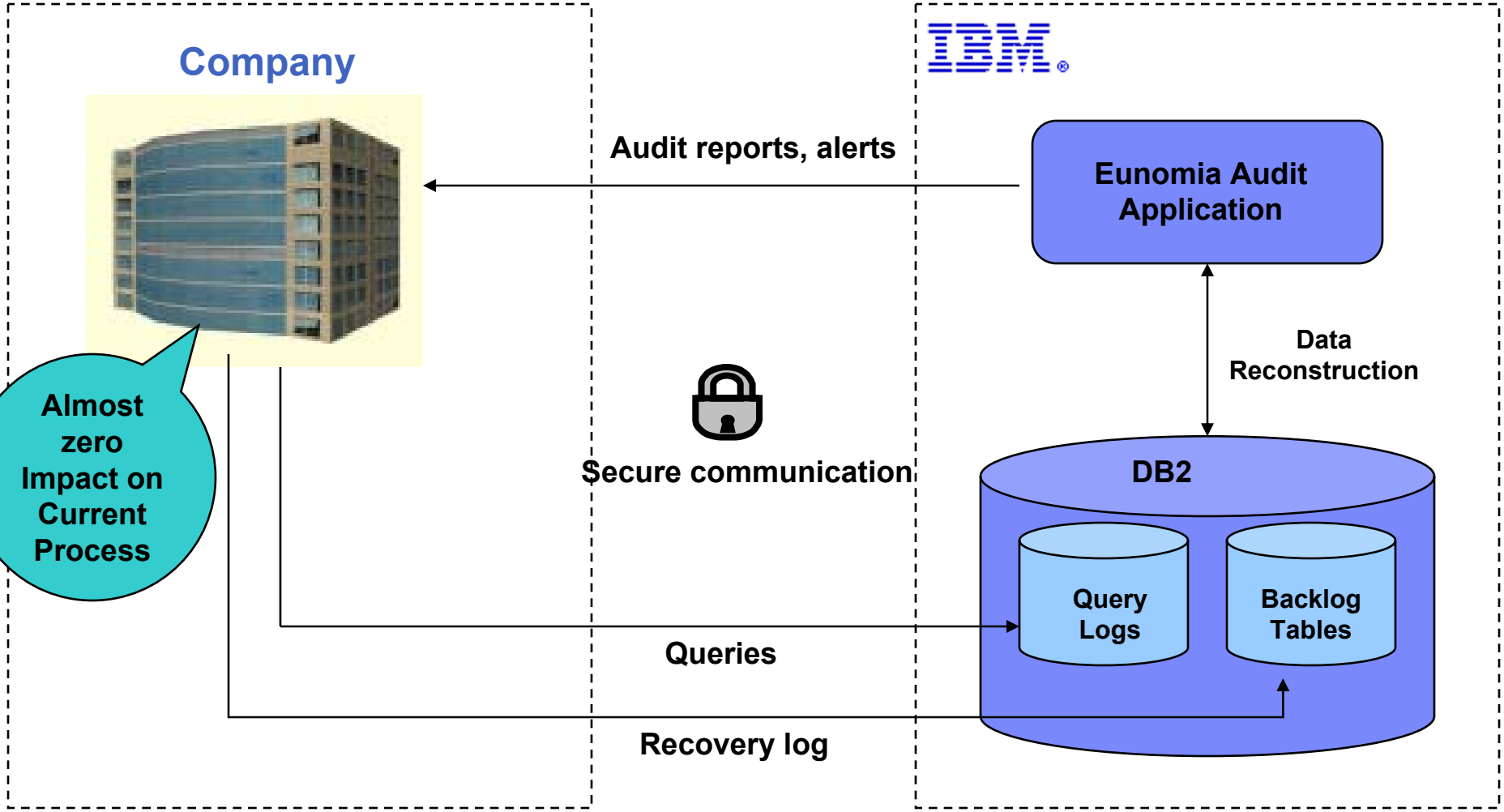
Eunomia In-House Model



- **Audit if some data has been disclosed in violation of the specified policy**
- **Audit queries specify what potential data disclosures need monitoring**
- **Identify logged queries which have accessed the compromised data**
- **Analyze circumstances of the violation**
- **Make necessary corrections to procedures, policies, security**

ID	Timestamp	Query	User	Purpose	Recipient
1	**2004-02…**	**Select …**	**Jane**	**Current**	**Ours**
2	**2004-02…**	**Select …**	**John**	**Telemarketing**	**public**

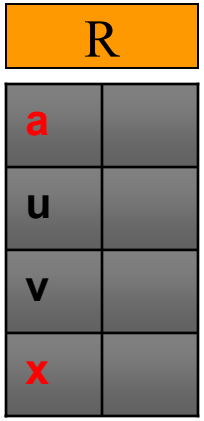
Eunomia Outsourced Model

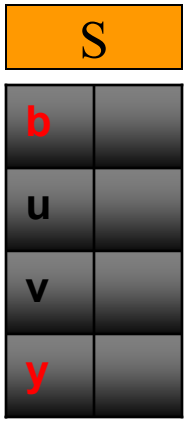


Sovereign Information Sharing

- Separate databases due to statutory, competitive, or security reasons.
  - ➢ Selective, minimal sharing on need-to-know basis.
- Example: Among those who took a particular drug, how many had adverse reaction and their DNA contains a specific sequence?
  - ➢ Researchers must not learn anything beyond counts.
- Algorithms for computing joins and join counts while revealing minimal additional information.

Minimal Necessary Sharing





R ❁ S

- ➢ R must not know that S has b & y
- ➢ S must not know that R has a & x

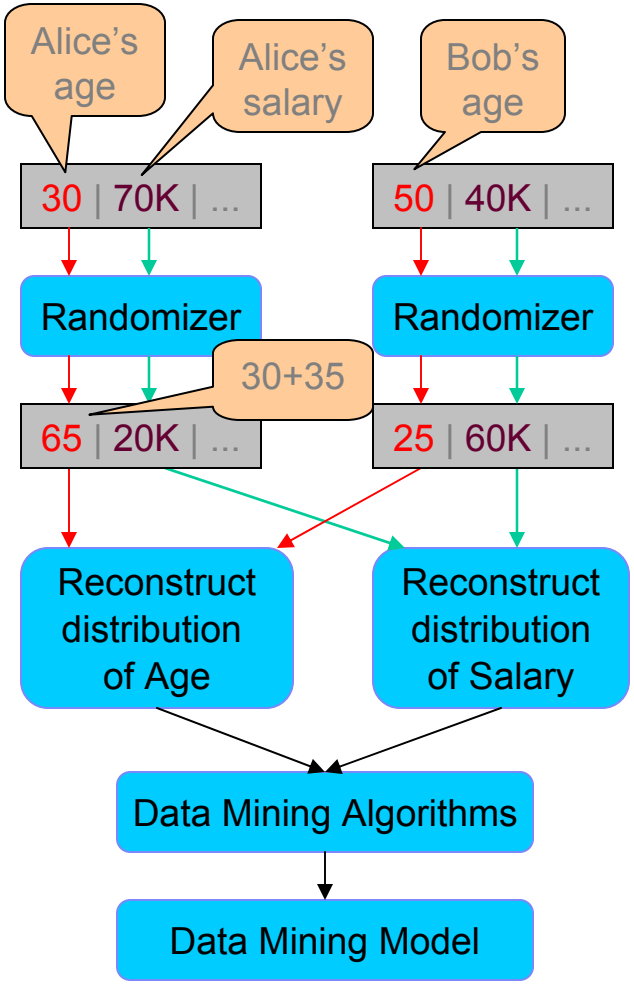
R ❁ S

u
v

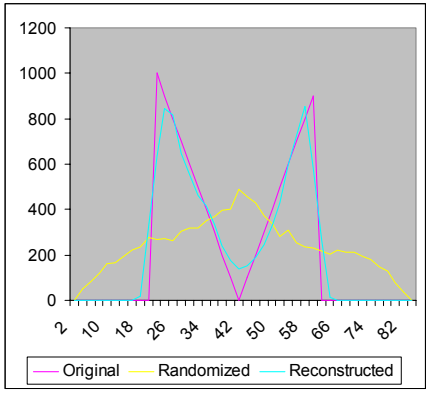
Count (R ❁ S)

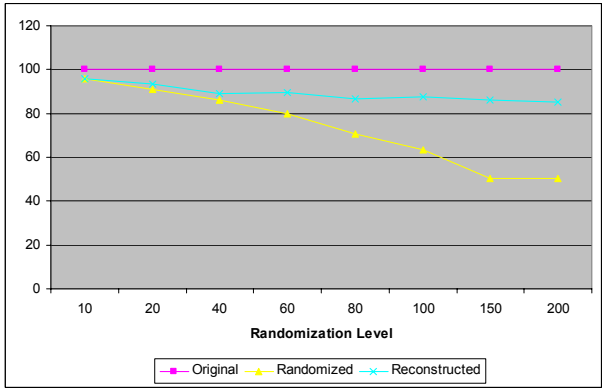
- ➢ R & S do not learn anything except that the result is 2.

Privacy Preserving Data Mining



- Insight: Preserve privacy at the individual level, while still building accurate data mining models at the aggregate level.
- Add random noise to individual values to protect privacy.
- EM algorithm to estimate original distribution of values given randomized values + randomization function.
- Algorithms for building classification models and discovering association rules on top of privacy-preserved data with only small loss of accuracy.





Conclusion

- **Data Disclosure will have to address Security and Privacy concerns for Grid applications**
- **There is technology that allows this to be done without impeding information flow.**

Technical Papers

- R. Agrawal, P. Bird, T. Grandison, J. Kieman, S. Logan, W. Rjaibi. Extending Relational Database Systems to Automatically Enforce Privacy Policies. Proc. of the 21st Int'l Conf. on Data Engineering (ICDE 2005), Tokyo, Japan, April 2005.
- R. Bayardo, R. Agrawal. Data Privacy Through Optimal k-Anonymization. Proc. of the 21st Int'l Conf. on Data Engineering (ICDE 2005), Tokyo, Japan, April 2005.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. Order Preserving Encryption of Numeric Data. ACM Int'l Conf. On Management of Data (SIGMOD), Paris, France, June 2004.
- R. Agrawal, A. Evfimievski, R. Srikant. Information Sharing Across Private Databases. ACM Int'l Conf. On Management of Data (SIGMOD), San Diego, California, June 2003.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. An Xpath Based Preference Language for P3P. 12th Int'l World Wide Web Conf. (WWW), Budapest, Hungary, May 2003.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. Implementing P3P Using Database Technology. 19th Int'l Conf.on Data Engineering(ICDE), Bangalore, India, March 2003.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. Server Centric P3P. W3C Workshop on the Future of P3P, Dulles, Virginia, Nov. 2002.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. Hippocratic Databases. 28th Int'l Conf. on Very Large Databases (VLDB), Hong Kong, August 2002.
- R. Agrawal, J. Kiernan. Watermarking Relational Databases. 28th Int'l Conf. on Very Large Databases (VLDB), Hong Kong, August 2002.
- A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke. Mining Association Rules Over Privacy Preserving Data. 8th Int'l Conf. on Knowledge Discovery in Databases and Data Mining (KDD), Edmonton, Canada, July 2002.
- R. Agrawal, R. Srikant. Privacy Preserving Data Mining. ACM Int'l Conf. On Management of Data (SIGMOD), Dallas, Texas, May 2000.

http://www.almaden.ibm.com/software/quest/Publications/ByDate.html