



Privacy in eHealth:

Leveraging Hippocratic Database Technology

Dr Tyrone W A Grandison
Manager, Intelligent Information Systems,
Healthcare Informatics, IBM Almaden Research Center,
San Jose, California

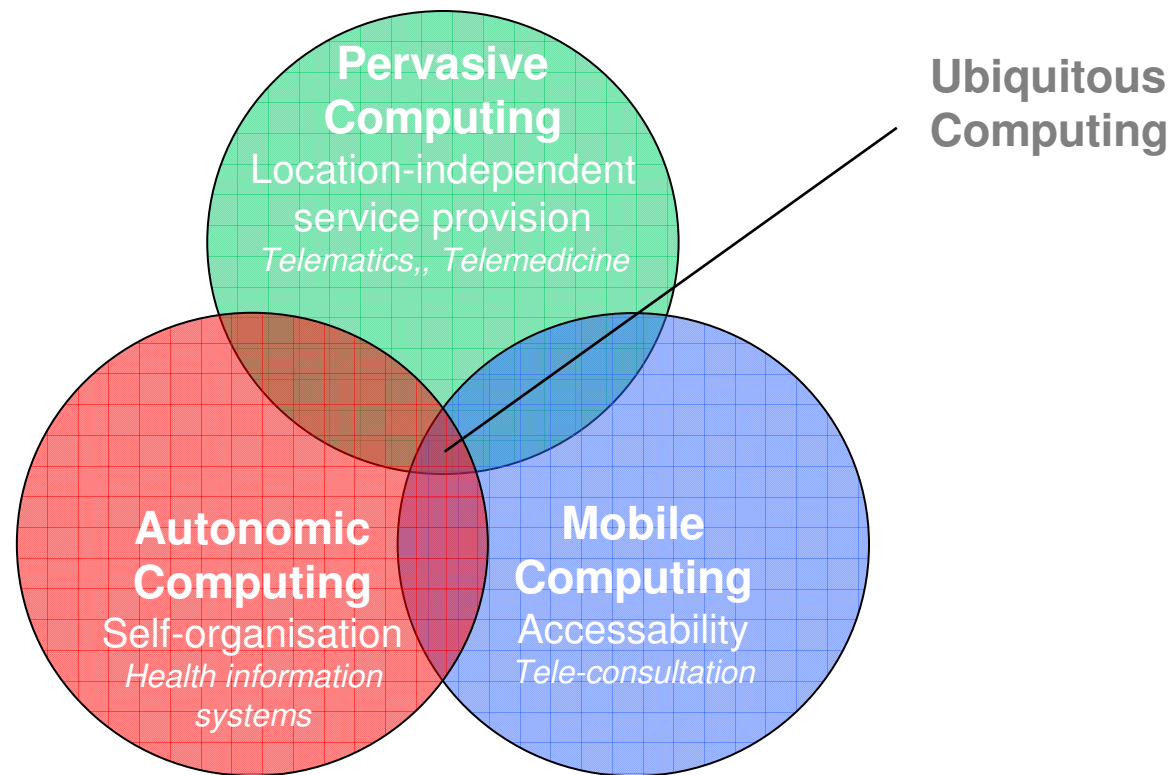
Information Security Institute
Queensland University of Technology
Guest Seminars on eHealth
Aug 23, 2007

Why Should I Worry About Privacy?

- Compliance with Data Protection Laws
 - Health Insurance Portability and Accountability Act
 - EU Data Protection Directive
 - Privacy laws in Canada, Japan, Australia
 - Gramm-Leach Bliley Act
- High profile privacy breaches and identity theft cases
 - A chronology of data breached from the Privacy Rights Clearinghouse noted that the total number of records confirmed stolen, from ChoicePoint incident in Feb 2005 till Nov 22, 2006, is 97,215,732
- Customer pressure for increased privacy and security
 - Ponemon Institute/MSNBC 2006 study asked, “Who do you trust more to protect your privacy — government or private corporations?” 88% picked the third option, “neither.”

The Promise of eHealth

Enabling the provision of electronic healthcare services, in a totally distributed environment.



- Bernd Blobel, Head, German National eHealth Competence Center, University of Regensburg Medical Center, Regensburg, Germany

Invariants for eHealth

■ eHealth systems mandates

- Support for Legacy Systems
 - Tightly Coupled Complex Systems
 - Each Silo'ed System has its own Protection Mechanisms
 - Conflicting Priorities and Policies
 - New (and changing) Technology
- Paradigm Shift
- Separation of Application Function from Security and Privacy function.
- Do not impact the performance/efficiency of the currently running system.
- Enable the current (clinical) workflow and do not require it to change.

Overview of Hippocratic Database Technologies

GOAL

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

Active Enforcement

Cell-level enforcement of disclosure policies and data subject preferences

Compliance Auditing

Determine whether data has been accessed in violation of specified policies

Sovereign Information Integration

Selective, minimal sharing across autonomous data sources, without trusted third party

Privacy-Preserving Data Mining

Preserves privacy at individual level, allowing accurate data mining models at aggregate level

Optimal

k-anonymization

De-identifies records in a way that maintains truthful data but is not prone to data linkage attacks

Database Watermarking

Tracks origin of leaked data by tracing a hidden bit pattern embedded in the data

Hippocratic Database Technology

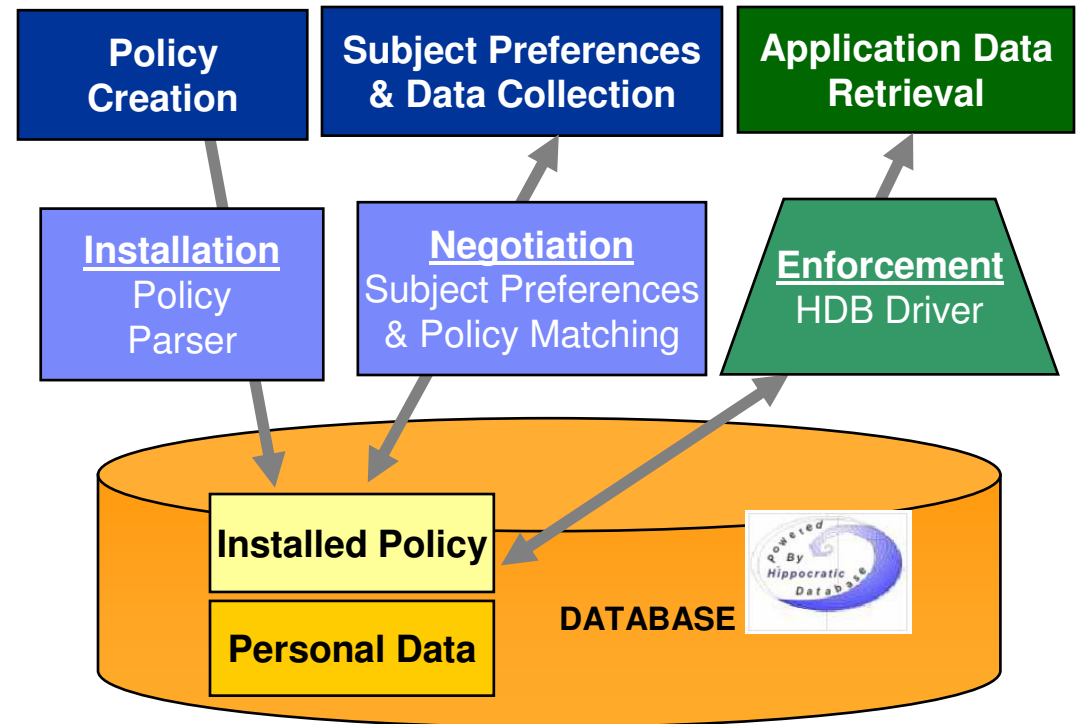
- Facilitates:
 - Automated,
 - Non-Intrusive,
 - High Performance,
 - Fine-Grained Data Disclosure,
 - At Database Level.

- Each component can be used alone or in conjunction with others, depending on the needs of the customer.

- Available Through IBM Channels: IBM Global Services (GBS), On-Demand Innovation Services (ODIS)

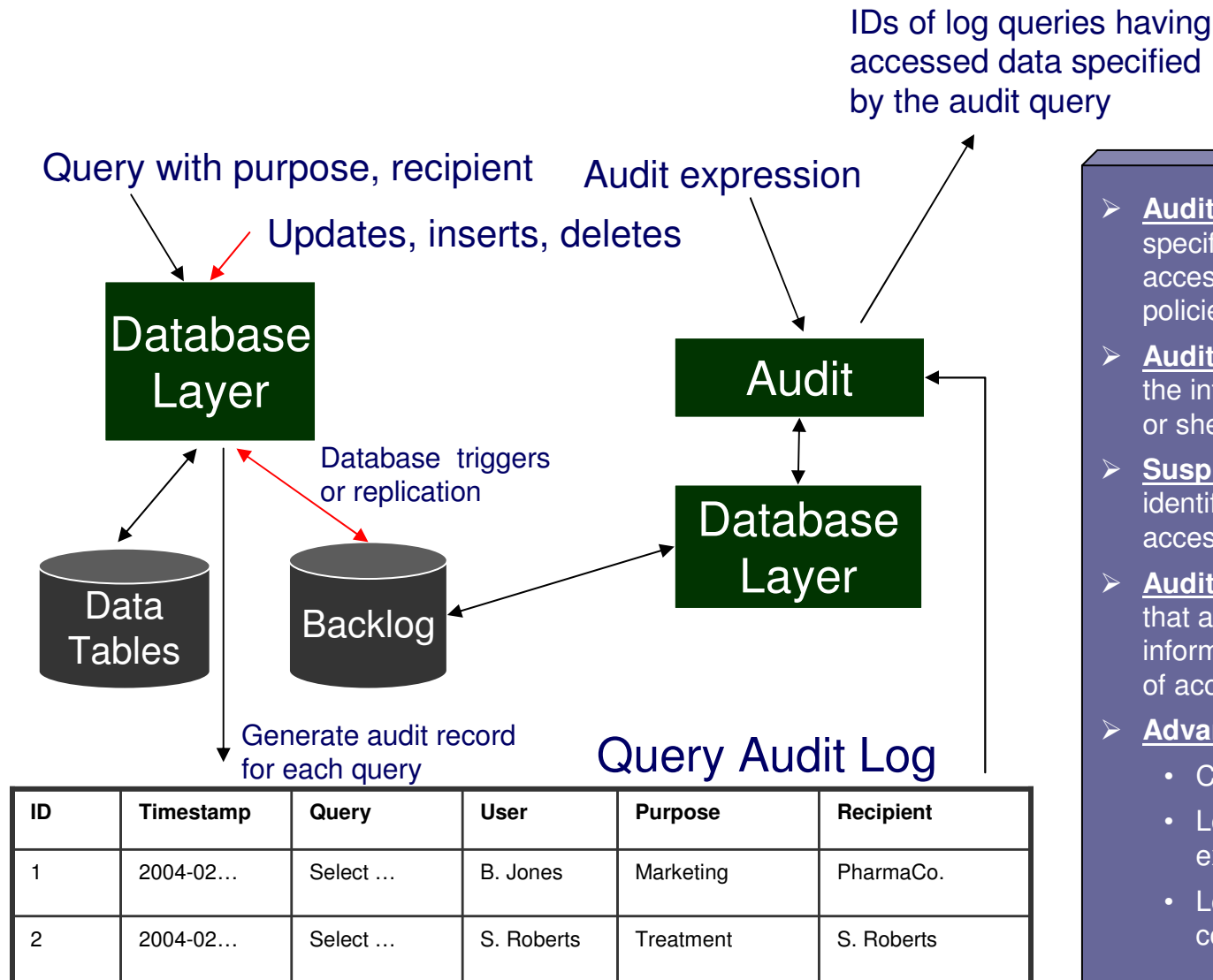
Hippocratic Database Active Enforcement

- **Privacy Policy:** Organizations define a set of policies describing who may access data (users or roles), for what purposes data may be accessed (purposes) and to whom data may be disclosed (recipients).
- **Consent:** Data subjects are given control, through opt-in and opt-out choices, over who may see their data and under what circumstances
- **Active Enforcement:** Intercepts and rewrites incoming queries to comply with policies, subject choices, and context.
- **Efficiency:** Rewritten queries benefit from all of the optimizations and performance enhancements provided by the underlying engine (e.g. parallelism).
- **Advantages:**
 - Cell-level access and disclosure control.
 - Application modification not required.
 - Database agnostic; does not require changes to the database engine.



#	Name	Age	Phone
1	Adam	25	(111) 111-1111
3	Bob	-	(333) 333-3333
4	Daniel	40	-

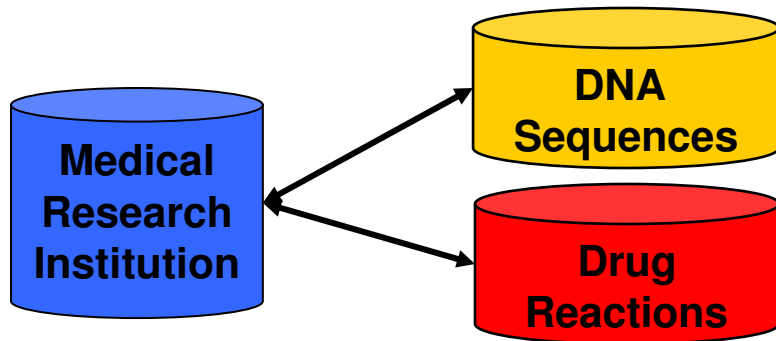
HDB Compliance Auditing



- **Audits:** Determine whether specified particular data has been accessed in violation of privacy policies or choices.
- **Audit expression:** Auditor specifies the information disclosures that he or she would like to track.
- **Suspicious Queries:** Audit system identifies logged queries that accessed the specified data
- **Audit Results:** Returns the queries that accessed the specified information and the circumstances of access.
- **Advantages:**
 - Cell-level disclosure auditing.
 - Low storage overhead; reuses existing database infrastructure.
 - Low performance impact; defers computation until audit time.

Sovereign Information Integration

- Autonomous databases for competitive, statutory, or security reasons.
 - Provides selective, minimal sharing on need-to-know basis.
- Example: Which DNA expressions correlate with reactions to certain drugs?
- Algorithms for computing secure joins and join counts without revealing any additional information among the databases.



Minimal Necessary Sharing

R	
a	
u	
v	
x	

S	
b	
u	
v	
y	

$R \bowtie S$

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

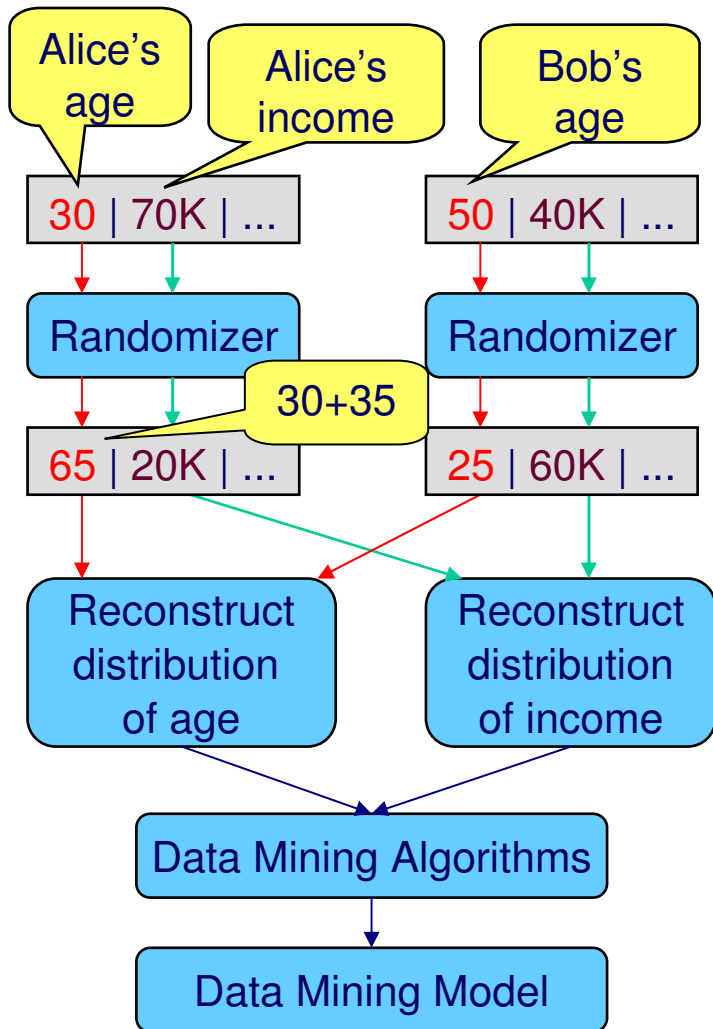
$R \bowtie S$

u
v

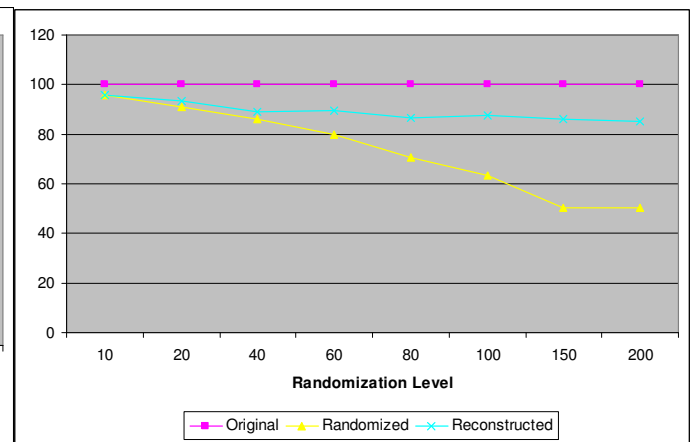
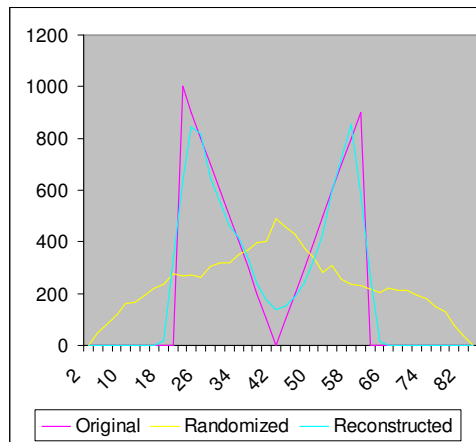
Count ($R \bowtie S$)

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

Privacy-Preserving Data Mining



- Preserves privacy at the individual level, but allows accurate data mining models to be constructed at the aggregate level.
- Adds random noise to individual values to protect data subject privacy.
- EM algorithm estimates 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.



Optimal k -Anonymization

- **Goal:** De-identify personal data such that it retains its integrity, but is resistant to data linkage attacks.
- **Motivation:** Naïve de-identification methods are prone to data linkage attacks, which combine subject data with publicly available information to re-identify represented individuals.
- **Samarati and Sweeney k -Anonymity* Method**
 - A k -anonymized data set has the property that each record is indistinguishable from at least $k-1$ other records within the data set.
- **Optimal k -Anonymization**
 - We have developed a k -anonymization algorithm that finds optimal k -anonymizations under two representative cost measures and variations of k .

Process of k -Anonymization

- **Data Suppression** - Involves deleting particular cell values or entire tuples.
- **Value Generalization** - Entails replacing specific values, such as a telephone number, with more general ones, such as the area code alone.

Advantages of Optimal k -anonymization

- **Truthful** - Unlike other disclosure protection techniques that use data scrambling, swapping, or adding noise, all information within a k -anonymized dataset is truthful.
- **Secure** - More secure than other de-identification methods, which may inadvertently reveal confidential information.

Name	Address	City	Age	Income
Erica	19 Main Street	San Jose	26	\$42,000
Paul	130 Harry Road	San Jose	42	\$88,000
Mark	4800 17th Street	San Jose	47	\$120,000
Henry	210 Almaden Pkwy	San Jose	28	\$50,000

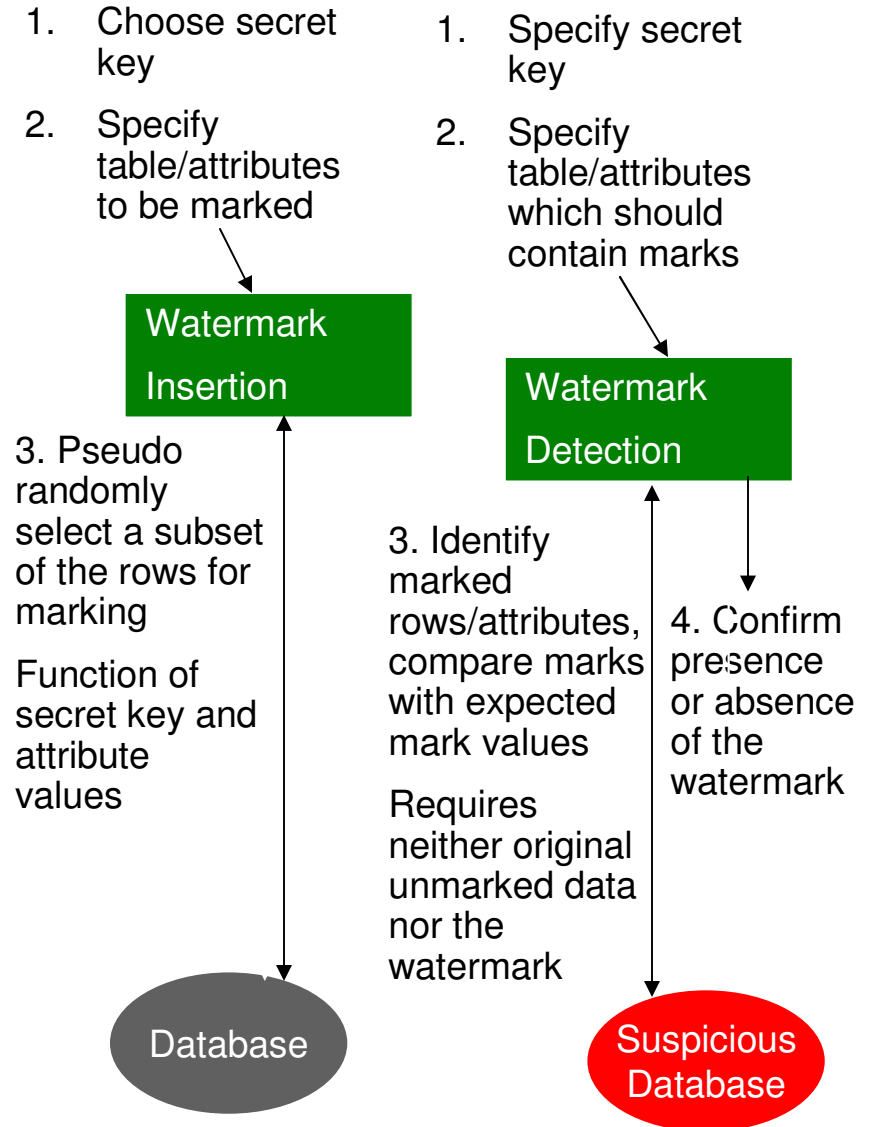
→
($k=2$, on name, address, age)

Name	Address	City	Age	Income
*	95131	San Jose	20-29	\$42,000
*	95120	San Jose	40-49	\$88,000
*	95120	San Jose	40-49	\$120,000
*	95131	San Jose	20-29	\$50,000

* P. Samarati and L. Sweeney. "Generalizing Data to Provide Anonymity when Disclosing Information." In Proc. of the 17th ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 188, 1998.

Database Watermarking

- **Goal: Deter data theft and assert ownership of pirated copies.**
- **Watermark** – Intentionally introduced pattern in the data.
 - Very unlikely to occur by chance.
 - Hard to find => hard to destroy (robust against malicious attacks).
- **Existing watermarking techniques developed for multimedia are not applicable to DB tables.**
 - Rows in a table are unordered.
 - Rows can be inserted, updated, deleted.
 - Attributes can be added, dropped.
- **New algorithm for watermarking DB tables.**
 - Watermark can be detected using only a subset of the rows and attributes of a table.
 - Robust against updates, incrementally updatable.



Conclusion

- eHealth is ushering in a new era in healthcare information systems.
- The ramifications of this new mode of operation have to be carefully examined.
- Privacy and security solutions require special consideration.
- There is room for terrific research and innovation, given that we abide by the invariants.



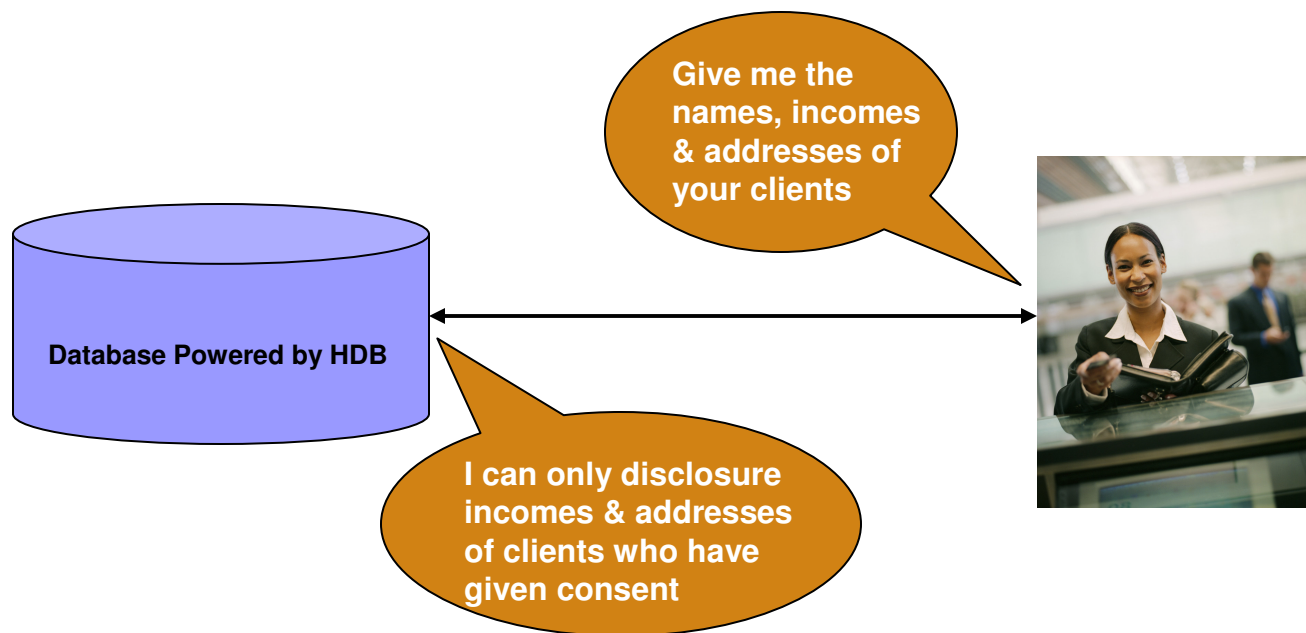
More Information: <http://www.almaden.ibm.com/software/disciplines/iis/>

Contacts: Dr Tyrone Grandison (tyroneg@us.ibm.com)

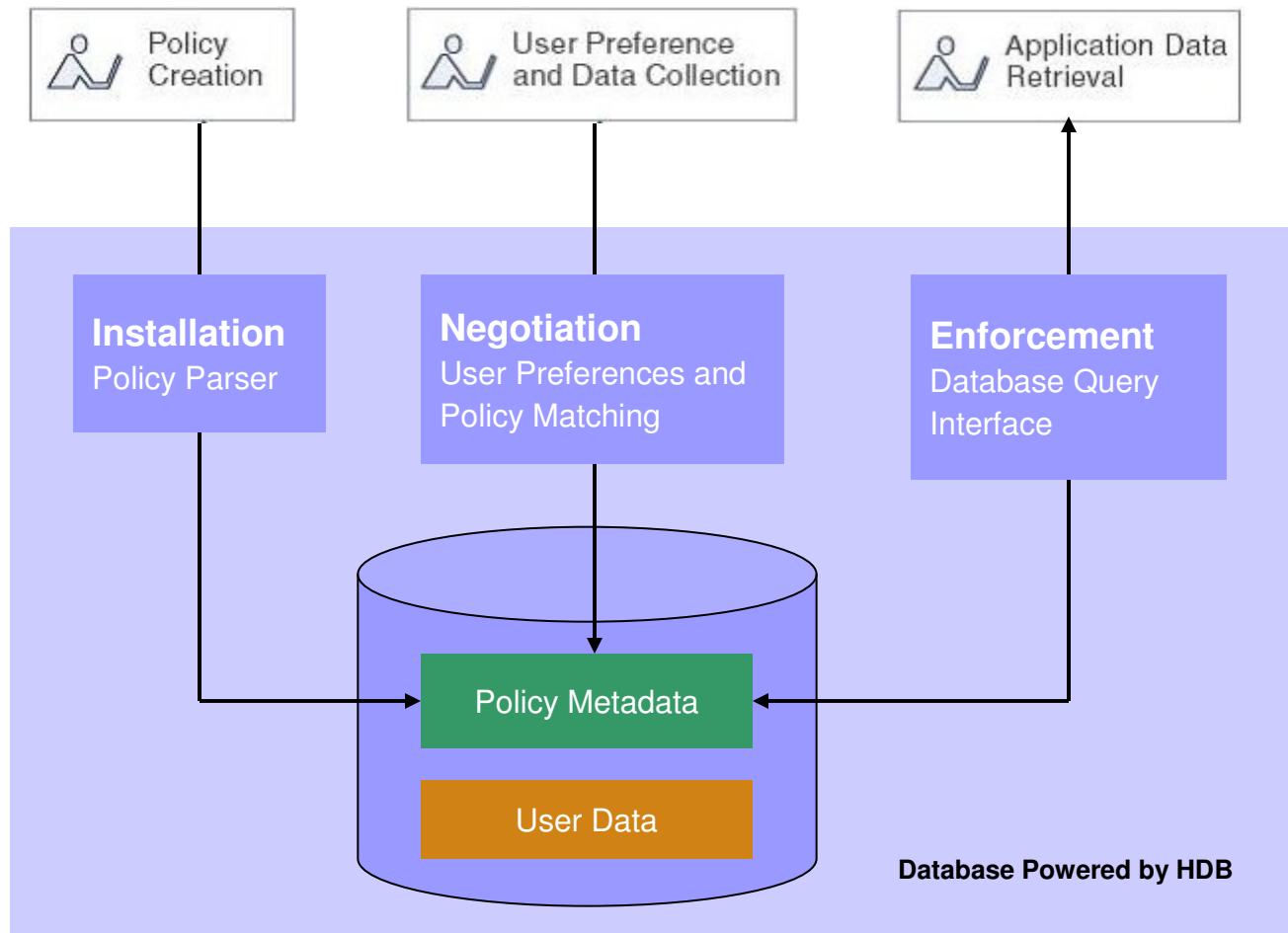
Backup Slides

<http://www.almaden.ibm.com/software/disciplines/iis/>

HDB Active Enforcement



HDB Active Enforcement

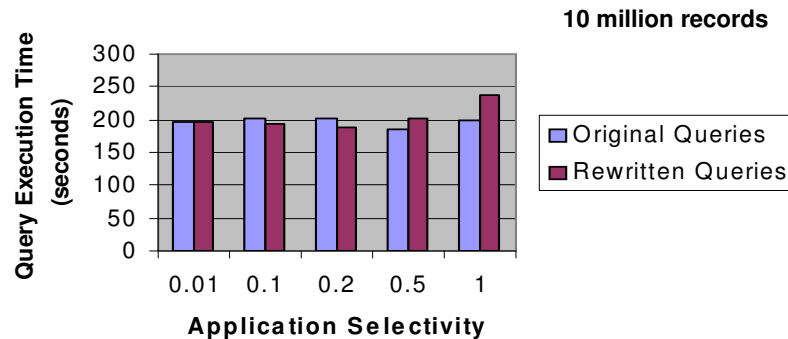


Enforcement: Value Proposition

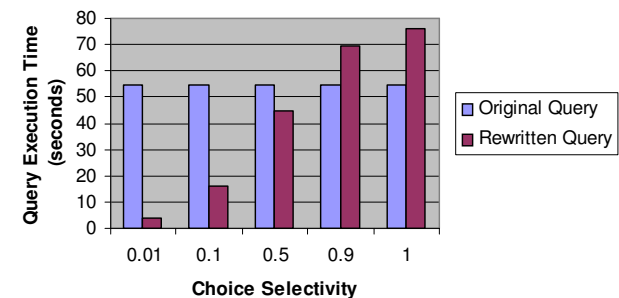
- Easy of Integration
 - Implementation intercepts and rewrites incoming queries to factor in policy, user choices, and context (e.g. purpose).
- Fine-Grained
 - Database-enforced disclosure control at cell-level of an organization's data policy and user preferences.
- Easier Enforcement after Policy Modification
 - Centralized and seamless policy creation and update.
- System Impact
 - Applications do not require any modification.

Enforcement: Value Proposition: cont'd

- Database agnostic
 - Does not require any change in the database engine.
- Reuses current features
 - Rewritten queries benefit from all the optimizations and performance enhancements provided by underlying engine (e.g. parallelism).
- Performance



Worst Case: Choice Selectivity = 1.
Everyone discloses everything. Query processing yields no value. The penalty is 5-15% of the execution time of the original query.



Standard Cases: Choice Selectivity varies.
In best case, HDB Active Enforcement gives an order of magnitude improvement.

HDB Active Enforcement Core Cell-Level Policy Enforcement

Example Scenario

ID	NAME	PHONE	SALARY
1	Alice	111-1111	10,000
2	Bob	222-2222	20,000
3	Carl	333-3333	30,000
4	David	444-4444	40,000

ID	PhoneChoice	SalaryChoice
1	0	1
2	1	0
3	0	0
4	1	1

For a certain user (data accessor) and purpose, **name** is allowed under the privacy policy, **phone** and **salary** are allowed on an opt-in basis.

HDB Active Enforcement Core Cell-Level Policy Enforcement : cont'd

```
SELECT Name, Phone, Salary  
FROM Customer
```

Results of query...

NAME	PHONE	SALARY
Alice	-	10,000
Bob	222-2222	-
Carl	-	-
David	444-4444	40,000

- Forbidden values covered by null values in resulting tables

HDB Compliance Auditing

Tell me who read W. Gates' financial and insurance information in 1987.



Compliance Auditing – Present Day

■ Concerns:

- Existing database systems and tools provide only offer 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.

■ Needed:

- 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.

HDB Compliance Auditing

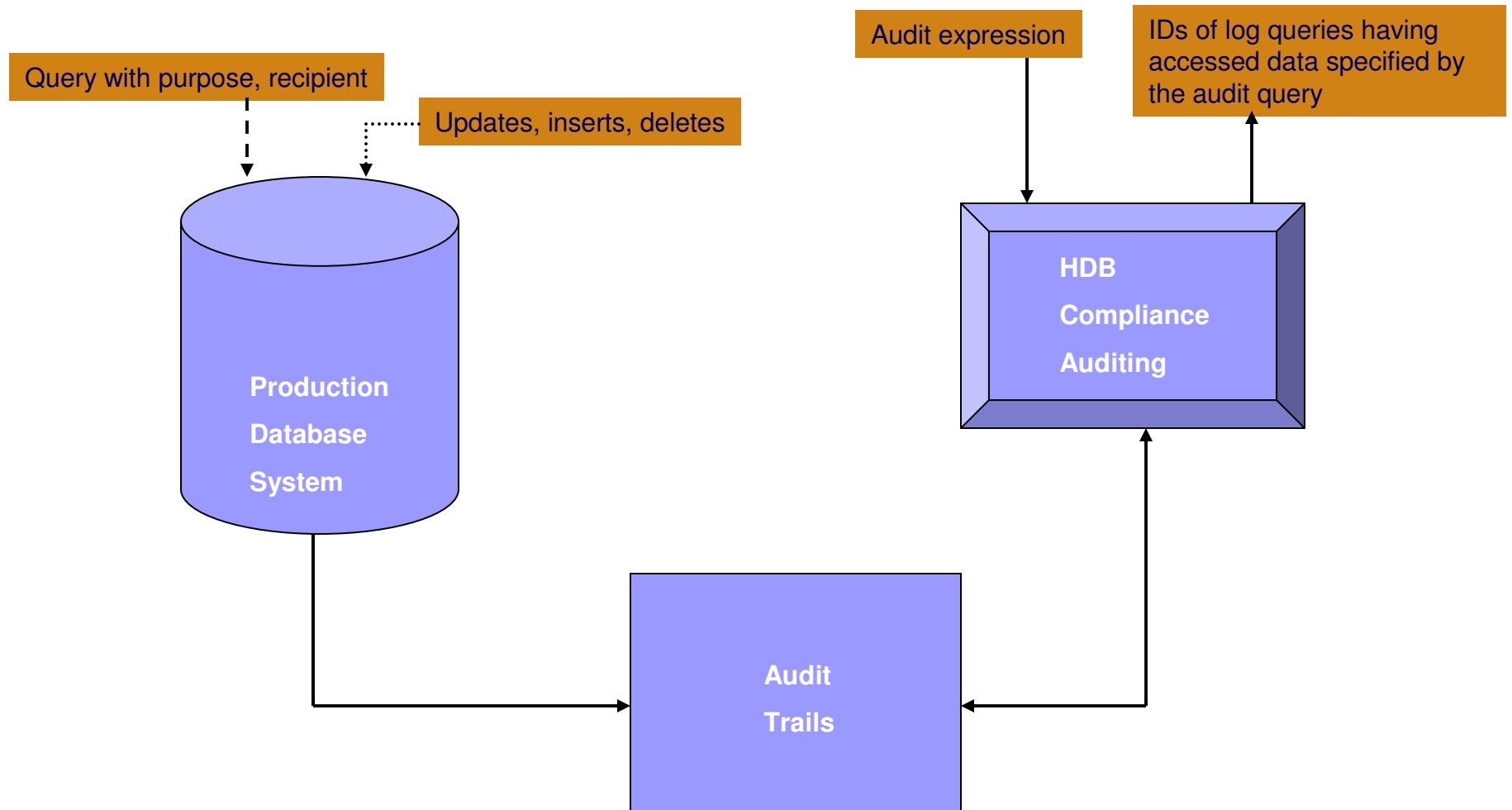
■ Infrastructure for Impact Minimization

- **Backlog table** can be populated with the update information by using database triggers or replication
- **Query logs** store id, timestamp, query, user & context (e.g. purpose & recipient)
- Backlog and query log generation significantly reduce storage and performance impact on production system. For zero impact, generation may be synchronized with routine backup activity.

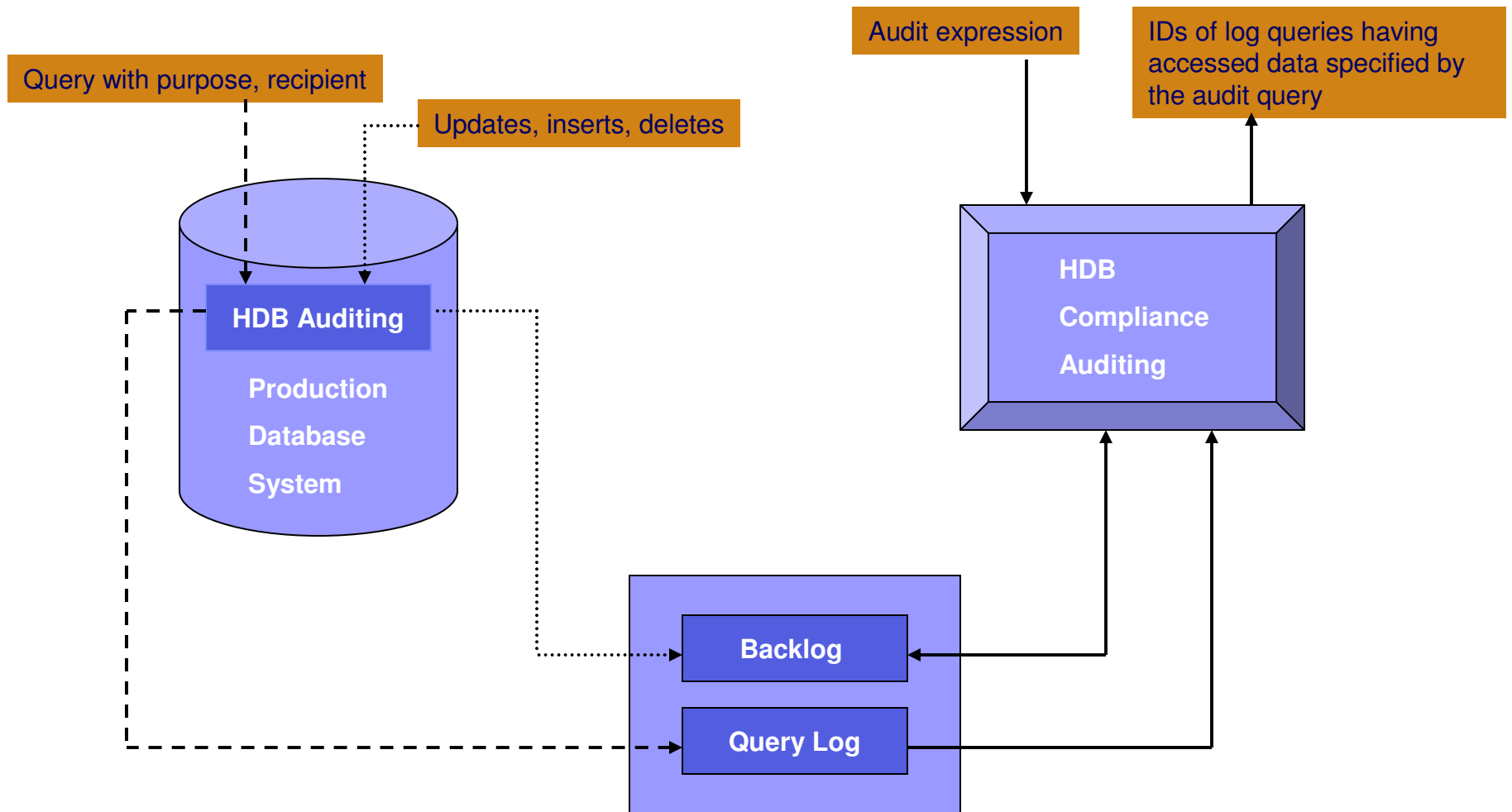
■ Functionality

- Audits whether particular data has been disclosed in violation of the specified policies
- Audit expression specifies what potential data disclosures need monitoring
- Identifies logged queries that accessed the specified data
- Analyze circumstances of the violation
- Make necessary corrections to procedures, policies, security

System Overview



System Overview: Detailed



Audit Scenario

- Claire is a customer of Astor Bank.
- She completes an application for a platinum card, providing Astor with current information about her employment, income, and assets.
- In notifying Astor of her privacy preferences, Claire opts out of disclosures of her financial information to unaffiliated third parties.
- After her application is approved, Claire receives several mailings from MortgageCo. at her office suggesting that she refinance her home. The interest rate offered is only available to those with incomes over \$100K.
- Claire then complains to Astor that it has disclosed her private financial information in violation of its privacy policy and her opt-out choice.
- Astor must now reveal all access of Claire's information to determine whether it was improperly disclosed to MortgageCo.

Audit Expression

Who has accessed Claire's income information?

audit	C.income
from	Customer C
where	C.name = 'Claire'

Problem Statement

- Given
 - A log of queries executed over a database
 - An audit expression specifying sensitive data

- Precisely identify
 - Those queries that accessed the data specified by the audit expression

Definitions (Informal)

- “Candidate” query
 - Logged query that accesses all columns specified by the audit expression
- “Indispensable” tuple (for a query)
 - A tuple whose omission makes a difference to the result of a query
- “Suspicious” query
 - A candidate query that shares an indispensable tuple with the audit expression

Example:

Query Q : Addresses of customers with incomes over \$100,000
Audit A : Claire’s income

Claire’s tuple is indispensable for both; hence query Q is “suspicious” with respect to A

HDB Compliance Auditing UI

PACT: Database Technology for Legislative Compliance - Microsoft Internet Explorer

Address: http://9.1.67.7:9080/demo2/admin?nextPage=auditResults.jsp&action=actionSort&removeColumn=&sortColumn=User&disclosureIndexes=&addTable=nothing&columnTable=nothing&BIG_CHECK=on

PACT: Database Technology for Legislative Compliance IBM.

Navigation

- [Create New Audit](#)
- [Administrative Console](#)
- [Disconnect from Database](#)

Connection Information

Database: dbtest
 Schema: DBUSER
 Logging: Enabled

Audit Results

create report Columns: + - 100% cancel search

		Payable)				
<input checked="" type="checkbox"/>	Click to view query details.	Parker, Peter (Accounts Payable)	Payment	None		2000-07-08 00:00
<input checked="" type="checkbox"/>	Click to view query details.	Parker, Peter (Accounts Payable)	Payment	None		2004-09-11 00:00
<input checked="" type="checkbox"/>	Click to view query details.	Parker, Peter (Accounts Payable)	Payment	None		2003-04-05 00:00
<input type="checkbox"/>	Click to view	Richards,				2004-

Compliance Auditing: Value Proposition

- Cost Reduction
 - Ability to monitor compliance and execute on compliance questions more efficiently and cost-effectively.
- Low Impact
 - Zero to minimal impact on company's current data operations depending on their requirements.
- Extensible
 - Inter-operates HDB Active Enforcement and other compliance technologies.
 - Backlog tables enable development of valuable customer insight applications.
- Security
 - Resistant to predicate-based attacks that return nonsensical output.

References

- R. Bayardo, R. Agrawal. “Data Privacy Through Optimal k -Anonymization.” *Proc. of the 21st Int'l Conf. on Data Engineering*, Tokyo, Japan, April 2005.
- R. Agrawal, R. Bayardo, C. Faloutsos, J. Kiernan, R. Rantzaou, R. Srikant. “Auditing Compliance with a Hippocratic Database.” *30th Int'l Conf. on Very Large Databases (VLDB)*, Toronto, Canada, August 2004.
- K. LeFevre, R. Agrawal, V. Ercegovac, R. Ramakrishnan, Y. Xu, D. DeWitt. “Limiting Disclosure in Hippocratic Databases.” *30th Int'l Conf. on Very Large Databases (VLDB)*, Toronto, Canada, August 2004.
- 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. “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.