



Hippocratic Database Technology

Privacy, Security & Compliance Technology

Intelligent Information Systems
IBM Almaden

What is Hippocratic Database technology?

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

- Technology that allows:
 - For the creation of a new generation of information systems that protect the privacy, security and ownership of data without impeding the flow of information.

Why should you deploy this technology?

- Compliance with Data Protection Laws
 - Gramm-Leach Bliley Act
 - Health Insurance Portability and Accountability Act
 - EU Data Protection Directive
 - Privacy laws in Canada, Japan, Australia
 - Payment Card Industry Data Security Standards
 - Interagency Guidelines for Safeguarding Customer Information
 - Basel II operational controls, Sarbanes-Oxley internal controls

- High profile privacy breaches and identity theft cases

- Customer pressure for increased privacy and security

Hippocratic Database (HDB) Functionality

- Covers data operations
 - Seven components that address specific client problems
 - Each component can be used alone or in conjunction with others, depending on the needs of the customer.

- Available Through The Following Channels
 - IBM Global Services (IGS)
 - Software Asset Deployment for your organization
 - On-Demand Innovation Services (ODIS)
 - Joint Projects

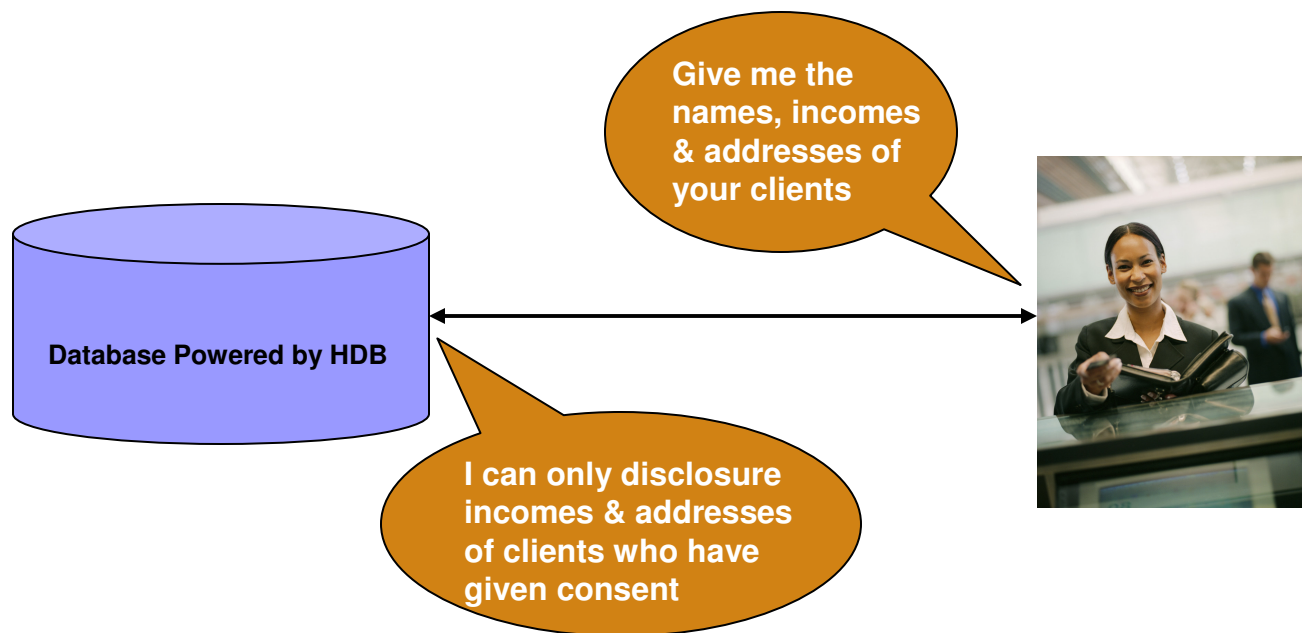
Functional Components:

- Active Enforcement
 - Enables the database to reveal only data compliant with policy.
- Compliance Auditing
 - Enables verification and monitoring of compliance with policy (e.g., legislation, privacy, security).
- Sovereign Information Integration
 - Enables two parties to securely and privately share common information without a third party involved.

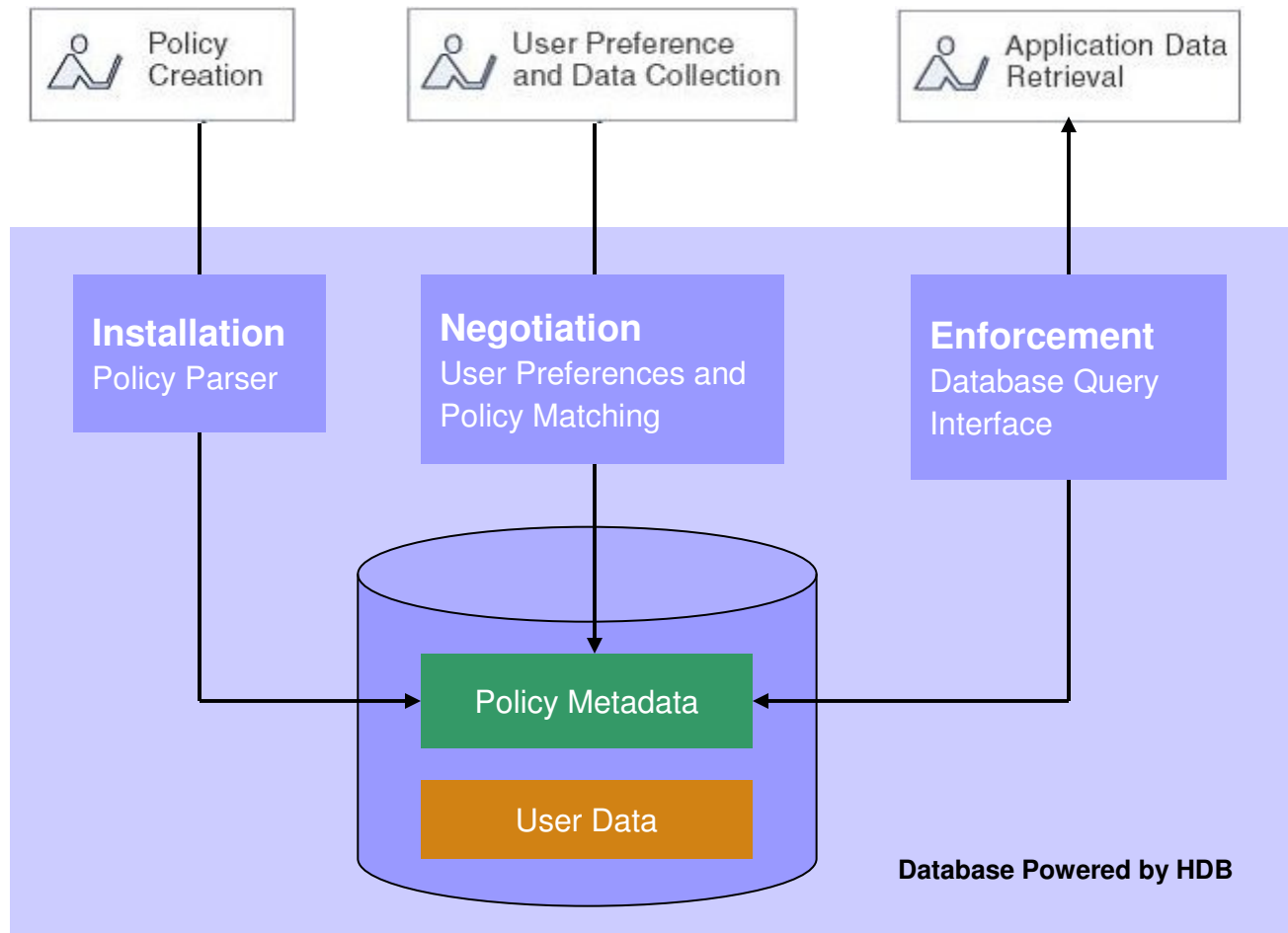
Functional Components: cont'd

- Order Preserving Encryption
 - Enables the protection of data from theft, while allowing encrypted data to be usable.
- Watermarking Databases
 - Deters data theft and asserts ownership of pirated copies
- Privacy Preserving Data Mining
 - Preserves privacy at the individual level, while enabling the construction of accurate data mining models at the aggregate level.
- BA k-anonymity
 - Enables data release, while preventing linking attacks and preserving data integrity

HDB Active Enforcement



HDB Active Enforcement

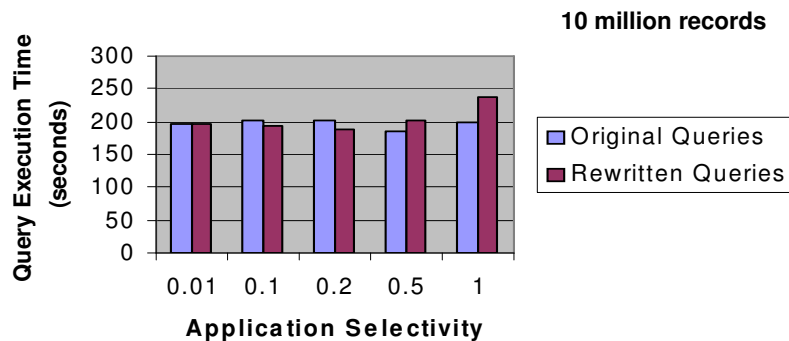


Enforcement: Value Proposition

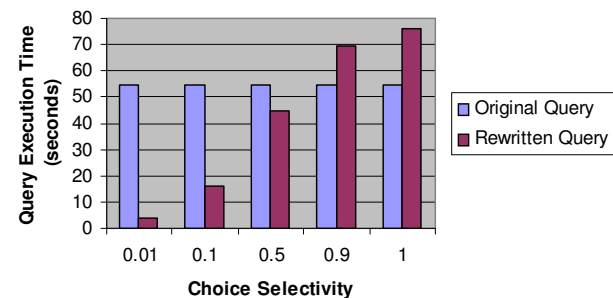
- Easy of Integration
 - Implementation intercepts and rewrites incoming queries to factor in policy, user choices, and context (e.g. purpose).
- Fine-Grained
 - Database-enforced disclosure control at cell-level of an organization's data policy and user preferences.
- Easier Enforcement after Policy 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

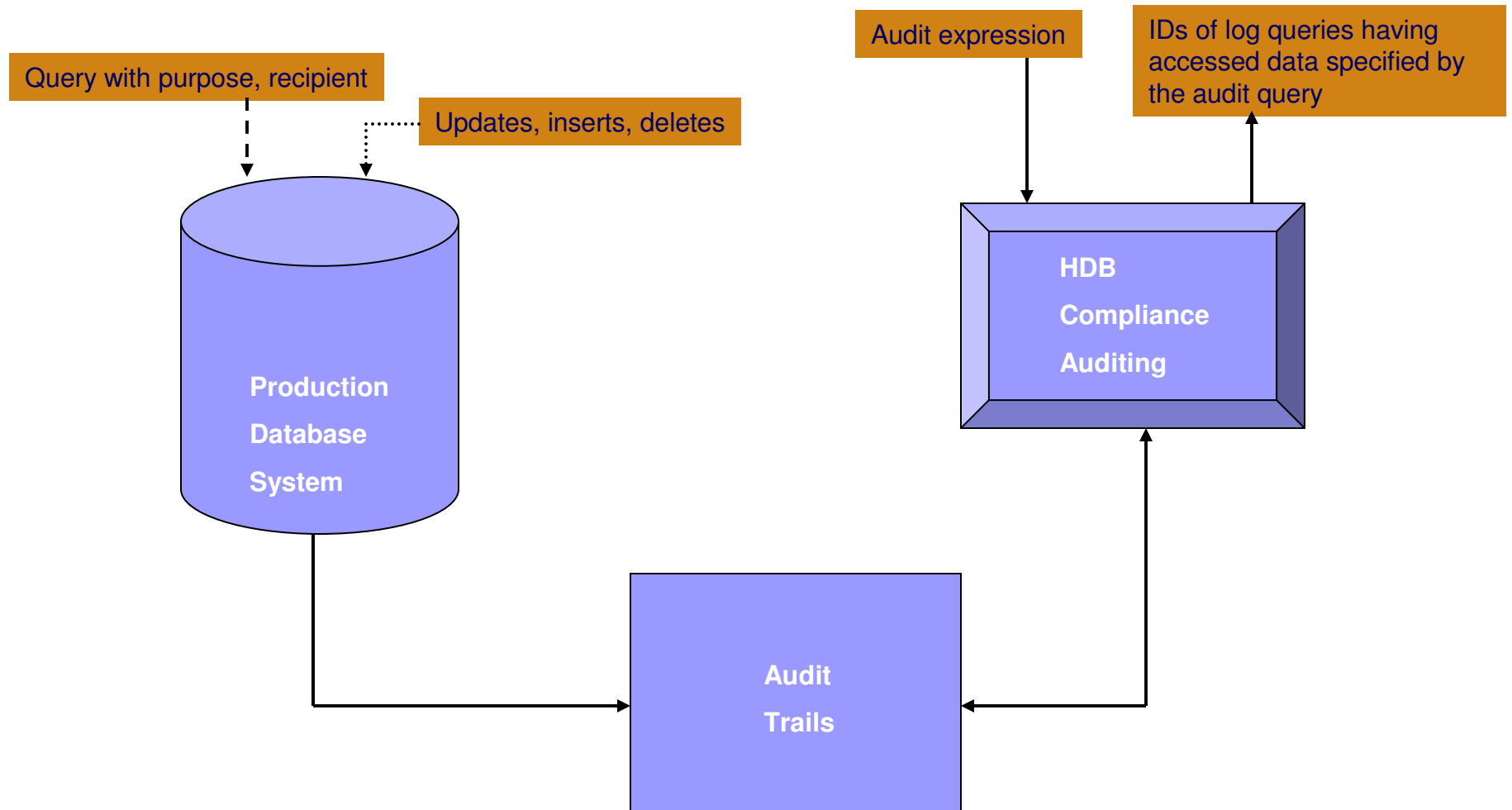
■ Infrastructure for Impact Minimization

- **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.

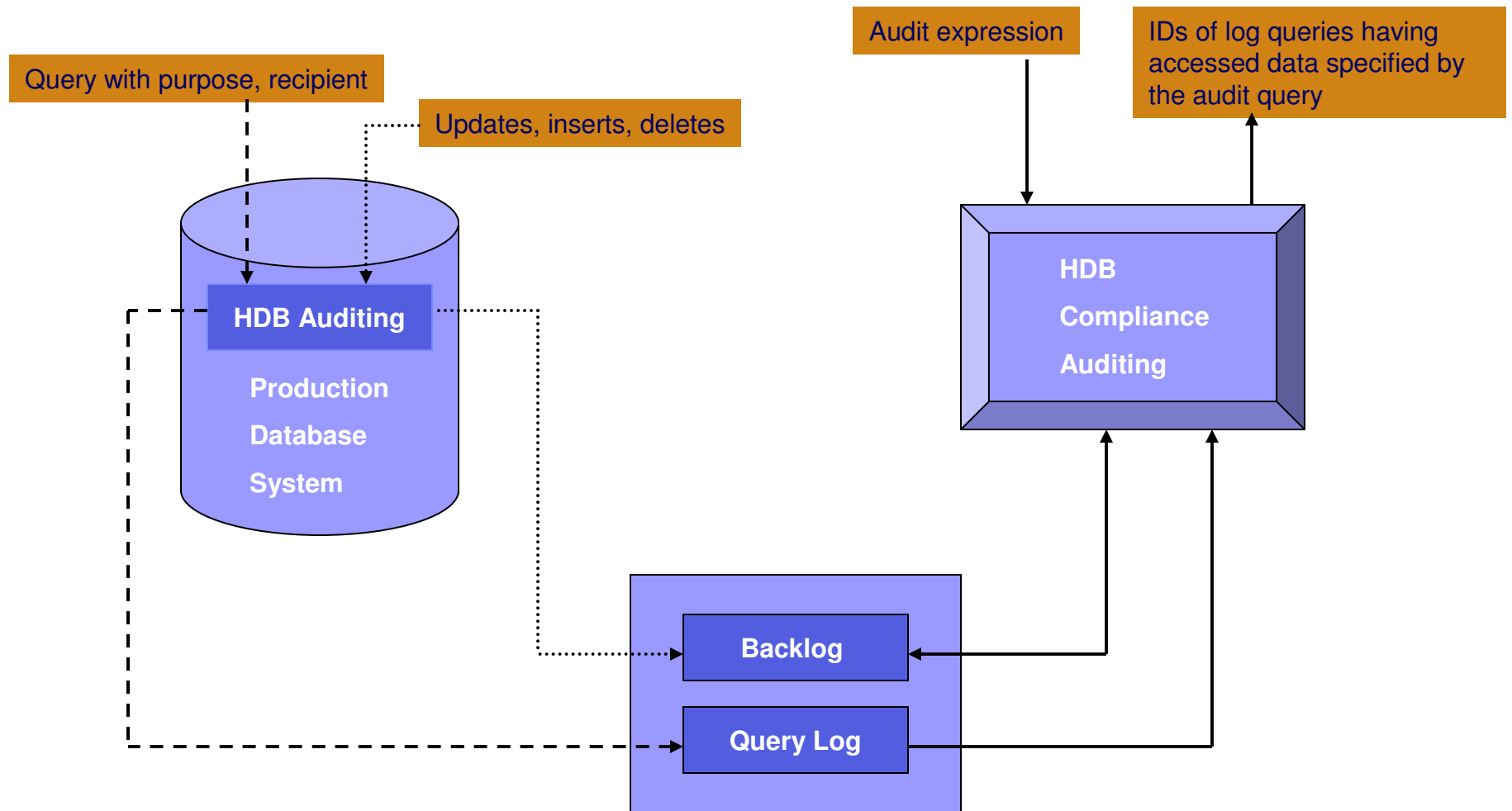
■ Functionality

- 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

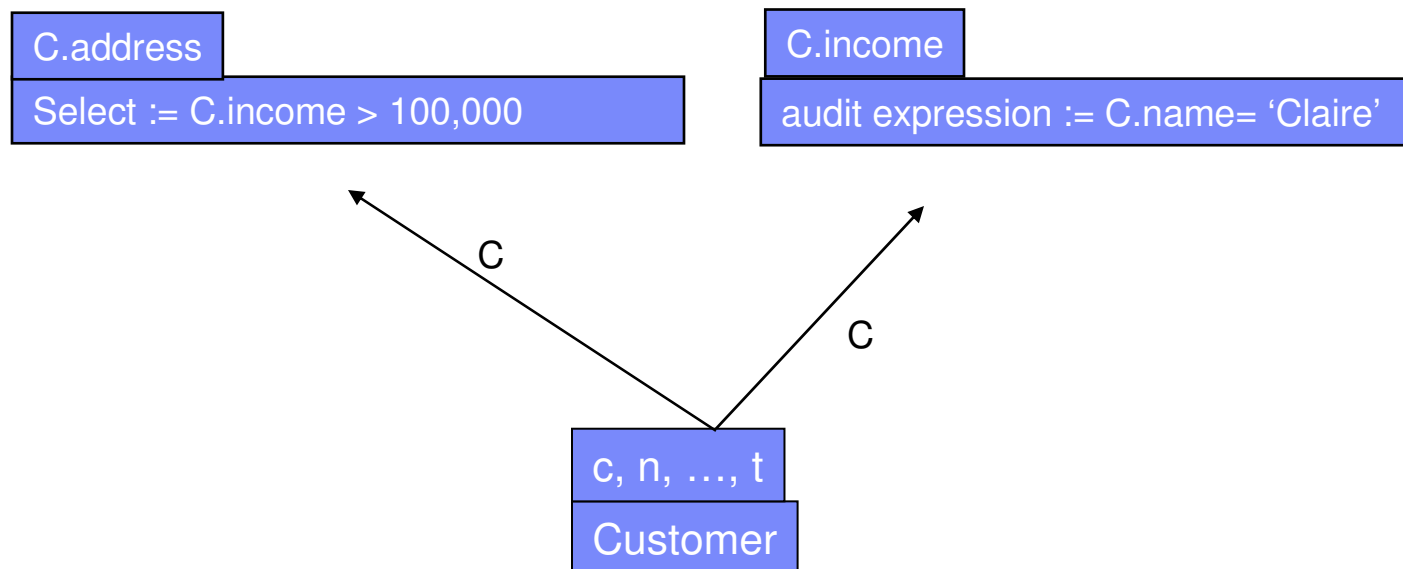
Example:

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

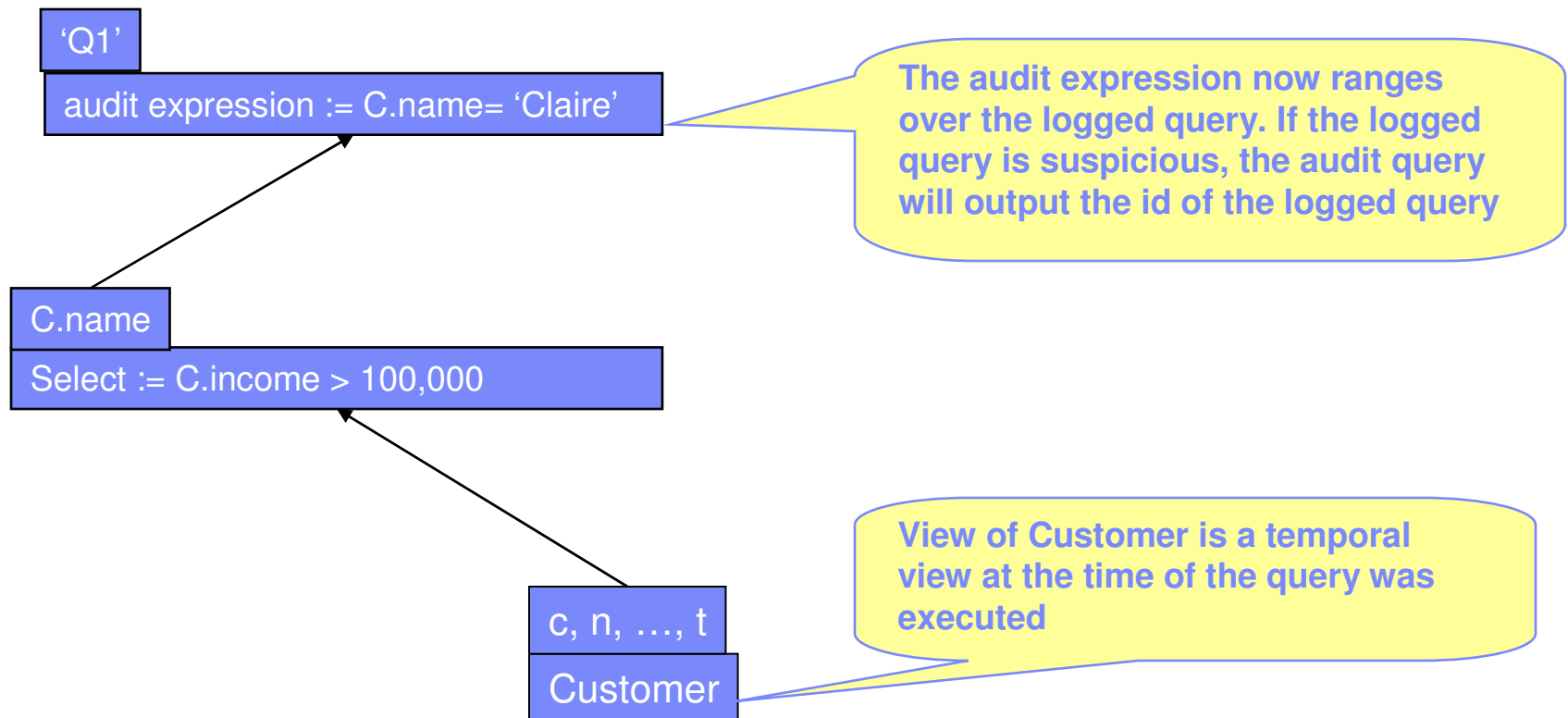
Claire’s tuple is indispensable for both; hence query Q is “suspicious” with respect to A

Merge Logged Queries and Audit Expression

Merge logged queries and audit expression into a single query graph



Transform Query Graph into an Audit Query



HDB Compliance Auditing UI

PACT: Database Technology for Legislative Compliance

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 checked="" type="checkbox"/>	Click to view query details.	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.



THE END

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

Backup Slides

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

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	
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 ($R \bowtie S$)

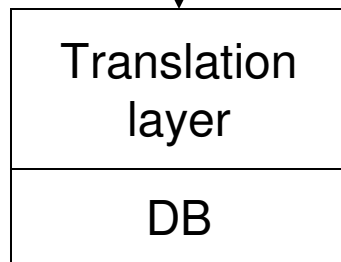
- R & S do not learn anything except that the result is 2.

Order Preserving Encryption (OPES)

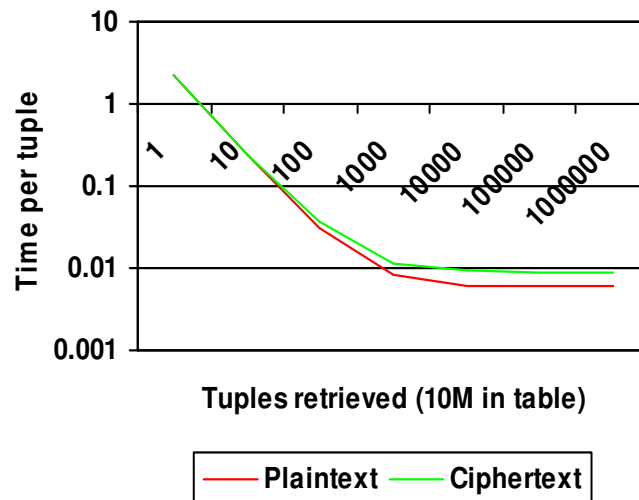


Plaintext Queries

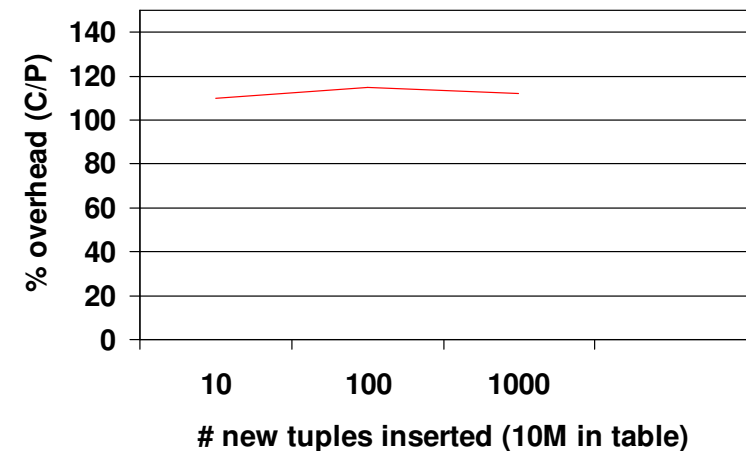
Select name from Emp where sal > 100000



Select decrypt ("xsxx", key1)
from "cwlxss"
Where
"xescs" >
OPESencr(100000, key2)

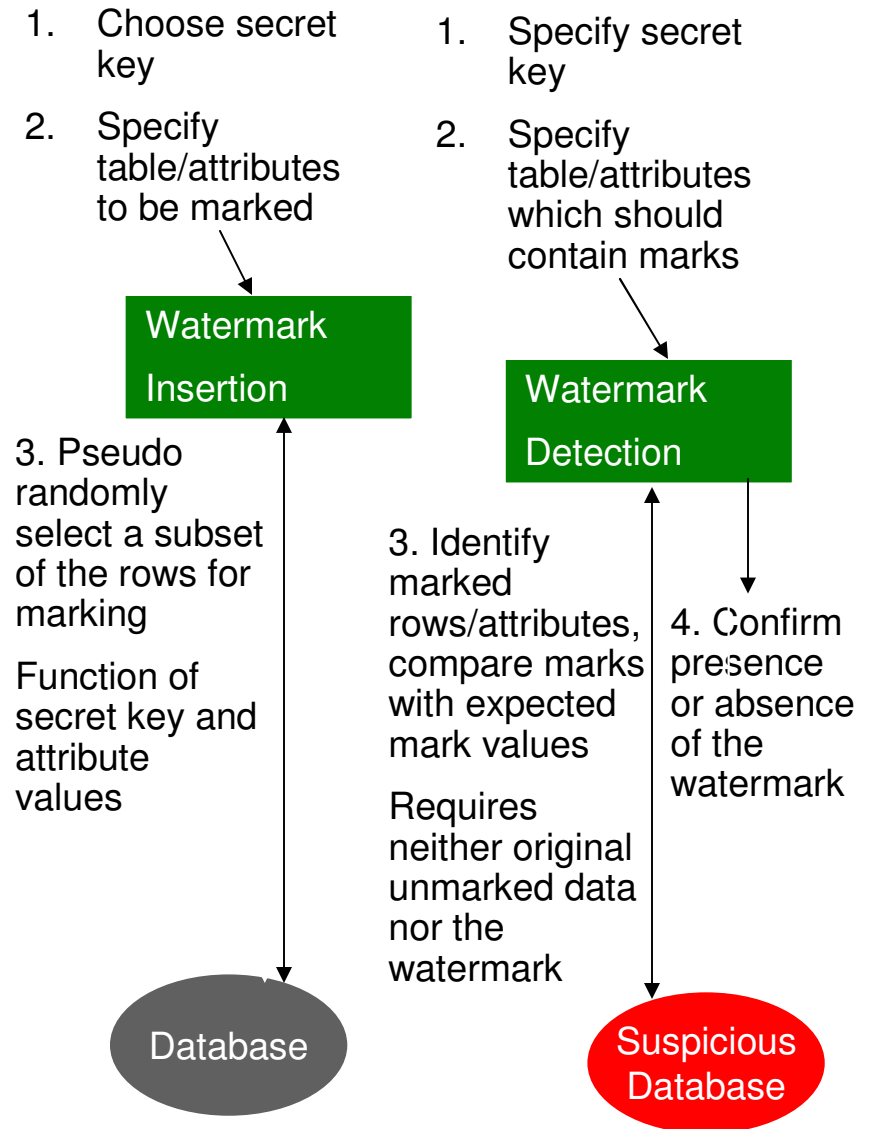


- Translation of plaintext queries into equivalent queries over encrypted data and metadata
- Use of regular as well as order preserving encryption for efficient evaluation of range queries over encrypted columns
- OPES encryption effectively hides the distribution of original plaintext values by encrypting input plaintext values into any chosen target distribution

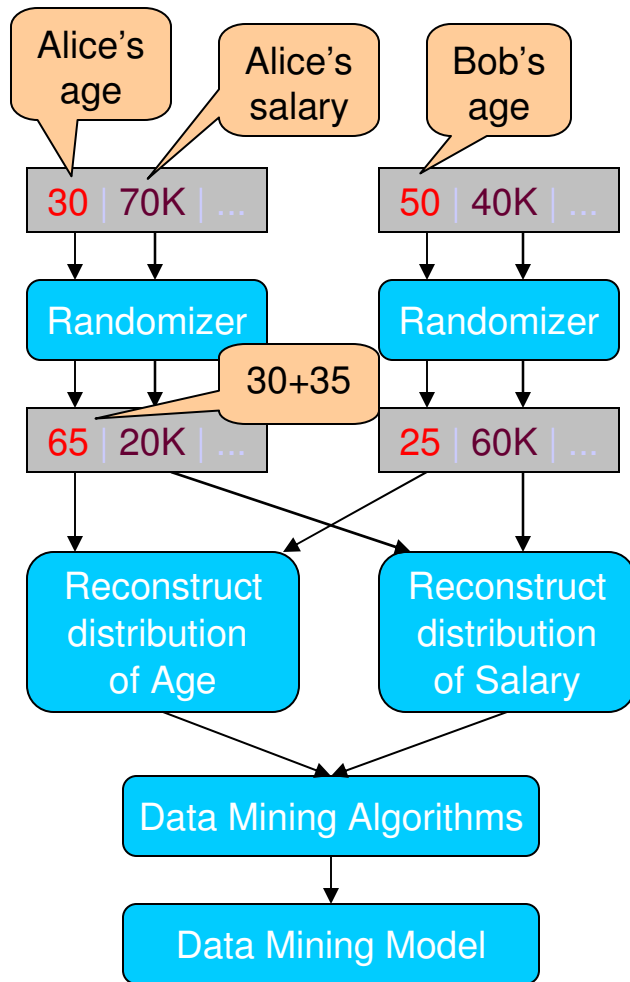


HDB Watermarking

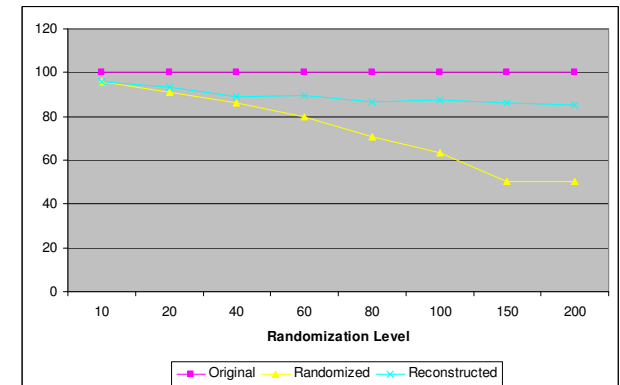
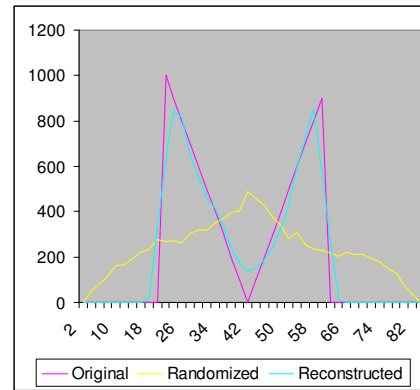
- **Goal: Deter data theft and assert ownership of pirated copies.**
- **Watermark – Intentionally introduced pattern in the data.**
 - Very unlikely to occur by chance.
 - Hard to find => hard to destroy (robust against malicious attacks).
- **Existing watermarking techniques developed for multimedia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



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.



Optimal *k*-Anonymization

- **Goal:** De-identify data such that it retains integrity, but is resistant to data linkage attacks.
- **Motivation:** Naïve methods are resistant to data linkage attacks, in 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 cell values or entire tuples.
- **Value generalization** - entails replacing specific values such as a phone number with a more general one, 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 17 th 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.

Sources

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

- White Papers
 - http://www.almaden.ibm.com/software/projects/iis/hdb/white_papers.shtml

- User Documents
 - http://www.almaden.ibm.com/software/projects/iis/hdb/user_docs.shtml

- Technical Papers
 - <http://www.almaden.ibm.com/software/projects/iis/hdb/publications.shtml>

References

- 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*, Tokyo, Japan, April 2005.
- R. Agrawal, R. Bayardo, C. Faloutsos, J. Kiernan, R. Rantzau, 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.