1

Overview of Data Anonymization

Points to Ponder

- What is data anonymization?
- What are the drivers for data anonymization?

Here are some startling statistics on security incidents and private data breaches:

- Leading technology and business research firms report that 70% of all security incidents and 80% of threats come from insiders and 65% are undetected.[1]
- *The Guardian* reports that a leading healthcare provider in Europe has suffered 899 personal data breach incidences between 2008–2011[2] and also reports that the biggest threat to its data security is its staff.[3]
- Datalossdb, a community research project aimed at documenting known and reported data loss incidents worldwide, reports that in 2011:
  - A major entertainment conglomerate found 77 million customer records had been compromised.[4]
  - A major Asian developer and media network had the personal information of 6.4 million users compromised.[4]
  - An international Asian bank had the personal information of 20,000 customers compromised.[4]

The growing incidence of misuse of personal data has resulted in a slew of data privacy protection regulations by various governments across countries. The primary examples of these regulations include the European Data Protection Directive and its local derivatives, the U.S. Patriot Act, and HIPAA.

Mischievous insiders selling confidential data of customer. (Courtesy of Jophy Joy)

The increasing trend of outsourcing software application development and testing to remote offshore locations has also increased the risk of misuse of sensitive data and has resulted in another set of regulations such as PIPEDA (introduced by the Canadian government).

These regulations mandate protection of sensitive data involving personally identifiable information (PII) and protected health information (PHI) from unauthorized personnel. Unauthorized personnel include the application developers, testers, and any other users not mandated by business to have access to these sensitive data.

The need to comply with these regulations along with the risk of hefty fines and potential loss of business in the event of misuse of personal data of customers, partners, and employees by insiders have led to enterprises looking at data privacy protection solutions such as anonymization. Data anonymization ensures that even if (anonymized) data are stolen, they cannot be used (misused)!!

PII

PII is any information which, by itself, or when combined with additional information, enables identification or inference of the individual. As a rule of thumb, any personally identifiable information that in the hands of a wrong person has the potential for loss of reputation or blackmail, should be protected as PII.

PII EXAMPLES

PII includes the following attributes.

Financial: Credit card number, CVV1, CVV2, account number, account balance, or credit balance

Employment related: Salary details

Personal: Photographs, iris scan, biometric details, national identification number such as SSN, national insurance number, tax identification number, date of birth, age, gender, marital status, religion, race, address, zip code, city, state, vehicle registration number, and driving license details

Educational details: such as qualifications, university course, school or college studied, year of passing

Contact information: including e-mail address, social networking login, telephone number (work, residential, mobile)

Medical information: Prior medical history/pre-existing diseases, patient identification number

PII DEFINITION

The National Institute of Standards and Technology (NIST) defines PII as any information that allows

- **Tracing of an individual or distinguishing of an individual:** This is the information which by itself identifies an individual. For example, national insurance number, SSN, date of birth, and so on.[5]

or

- **Linked or linkable information about the individual:** This is the information associated with the individual. For example, let's assume a scenario where the first name and educational details are stored in one data store, and the last name and educational details are in another data

> store. If the same individual can always access both data stores, this individual can link the information to identify another individual. This is a case of linked information. If the same individual cannot access both data stores at the same time, or needs to access both data stores separately, it is a case of linkable information.[5]
>
> Thus if both data stores do not have controls that allow for segregation of data stores, it is an example of linked information. If the data stores have segregating security controls, it is linkable information.

PHI

A lot of personal health information is collected, generated, stored, or transmitted by healthcare providers. This may be past health information, present health information, or future health information of an individual. Health may point toward physical or mental health or both. Such information directly or indirectly identifies the individual. The difference between PII and PHI is that PHI does not include education or employment attributes. The introduction of the Health Insurance Portability and Accountability Act (HIPAA) by the United States brought in the necessary urgency among organizations toward protection of PHI. PHI covers all forms of media (electronic, paper, etc.).

What Is Data Anonymization?

Data anonymization is the process of de-identifying sensitive data while preserving its format and data type.

The masked data can be realistic or a random sequence of data. Or the output of anonymization can be deterministic, that is, the same value every time. All these are dependent on the technique used for anonymization.

Technically, data masking refers to a technique that replaces the data with a special character whereas data anonymization or data obfuscation constitutes hiding of data and this would imply replacement of the original data value with a value preserving the format

and type. Thus, replacing "Don Quixote" with "Ron Edwards" would be a case of data anonymization whereas replacing "Don Quixote" with "XXXXXXXXXX" would be a case of data masking.

However, colloquially, data masking, data anonymization, data de-identification, and data obfuscation are interchangeably used and hence in this book, for all purposes, data anonymization and data masking are used interchangeably. In this book, when we are looking at data masking technically, "character masking technique" would be explicitly mentioned.

What Are the Drivers for Data Anonymization?

The need for data anonymization can be attributed to the following key drivers:

- The need to protect sensitive data generated as part of business
- Increasing instances of misuse of personal data and resultant privacy issues
- Astronomical cost to the business due to misuse of personal data
- Risks arising out of operational factors such as outsourcing and partner collaboration
- Legal and compliance requirements

The Need to Protect Sensitive Data Handled as Part of Business

Today's enterprises handle enormous amounts of sensitive data as part of their business. The sensitive data can be the personally identifiable information of customers collected as part of their interactions with them, the personally identifiable information of employees including salary details collected as part of their HR (Human Resource) processes, or protected health information of their customers and employees. Enterprises collect, store, and process these data and may need to exchange these data with their partners, outsourcing vendors. Misuse of any of this information poses a serious threat to their business. In addition to PII and PHI, enterprises also handle a lot of classified information that should not be made available to the public or to partners or to competitors and these also need to be protected from any misuse.

Increasing Instances of Insider Data Leakage, Misuse of Personal Data, and the Lure of Money for Mischievous Insiders

Based on its research from various cybercrime forums, a leading U.S. newspaper has found interesting statistics on the black market of private data. The study shows that leaking the driver's license information of one person can fetch between $100–$200, and billing data, SSN, date of birth, and credit card number can fetch a higher price.[6]

With such a booming black market for personally identifiable information, it is no wonder that the incidences of misuse of personal data by insiders have increased.

Misuse of personal data can be intentional or unintentional.

Employees Getting Even with Employers Monetary gain is not the sole motivator for misuse of personal data by insiders. Cases have come to light where dissatisfied employees or contractors have leaked or misused personal data of customers just to get back at the company or organization. This has resulted in a serious loss of image and business to these companies.

Employees getting even with employers. (Courtesy of Jophy Joy)

Negligence of Employees to Sensitivity of Personal Data Concerns related to loss of privacy of customers, partners, and employees have not arisen just due to intentional misuse of customers' personal data.

Employee or organizational negligence has also contributed to this. The need to appear helpful to those asking for personal information, lack of sensitivity when dealing with personal data, absence of information privacy policies, or lack of adherence to information privacy policies of companies due to minimal awareness have all contributed to the misuse of personal data. Despite privacy regulations being passed by various governments, we still see organizations using the personal data of customers collected for business purposes for marketing activities and many customers are still unaware of this.

Negligence of employees regarding sensitivity of personal data. (Courtesy of Jophy Joy)

Astronomical Cost to the Business Due to Misuse of Personal Data

In addition to loss of customer trust and resultant attrition, any misuse of personal data of customers or employees involves the need to engage lawyers for legal defense. Most cases of personal data misuse also end up in hefty fines for the enterprises thus making the cost extremely expensive.

In March 2011, Ponemon Institute, a privacy think tank, published their sixth annual study findings on the cost of data breaches to U.S.-based companies. This benchmark study was sponsored by Symantec and involved detailed research on the data breach experiences of more than 50 U.S. companies cutting across different industry sectors

Table 1.1 Estimated Cost of Data Breach[7]

DESCRIPTION	ESTIMATE
Cost of every compromised customer record per data breach incident	$214
Average total cost per incident	$7.2 Million
Average customer churn rates (loss of customers who were affected by the data breach incident after being notified of this breach)	4%

including healthcare, finance, retail, services, education, technology, manufacturing, transportation, hotels and leisure, entertainment, pharmaceuticals, communications, energy, and defense.

Each of the data breach incidents involved about 1,000 to 100,000 records being compromised. This study arrived at an estimate of the per-customer record cost and average per-incident cost, as well as the customer churn rate as a result of the breach. The figures are shown in Table 1.1.

The direct cost factors that have gone into the above estimate include expensive mechanisms for detection, escalation, notification, and response in addition to legal, investigative, and administrative expenses, customer support information hotlines, and credit monitoring subscriptions. The indirect or resultant cost factors include customer defections, opportunity loss, and reputation management.

Risks Arising out of Operational Factors Such as Outsourcing and Partner Collaboration

Outsourcing of IT application development, testing, and support activities result in data moving out of the organization's own premises as well as data being accessible to employees of the contracted organizations.

Collaboration with partners increasingly involves exchange of data. For example, a healthcare company would need to exchange patient data with the health insurance provider.

Thus outsourcing and partner collaboration increases the risk of misuse of personal data manifold.

Legal and Compliance Requirements

When governments and regulatory bodies get their act together, they bring in legislation that ensures the risk of litigation remains high for businesses. Businesses respond by turning back to their legal

department to ensure that they comply with the new regulations. Most governments have a "herd mentality" especially when it comes to issues that are global or have the potential to become global. When one friendly country passes legislation, it is just a matter of time before another country's government passes similar legislation.

This is what happened to "Protection of Data Privacy." The frequent incidents around identity theft and misuse of sensitive data ensured that the European Union passed the European Data Protection Directive and each of the countries belonging to the Union passed its own version of the European Data Protection Act. Meanwhile, the United States passed the "Patriot Act," the HIPAA, and Gramm–Leach–Bliley Act (GLBA), and Canada passed the PIPEDA act. All these acts focused on protection of sensitive personal data or protected health data.

Not to be left behind, the payment card industry came up with its own data security standards for protecting the consumer's credit card information. This set of standards was called "PCI-DSS" and imposed hefty fines on retailers and financial institutions in case of a data breach related to consumer credit cards. However, they also incentivize the retailers and financial institutions for adopting PCI-DSS (and showing evidence of this). Implementation of PCI-DSS on their IT systems lessens the probability of leakage or misuse of consumer credit card information. An overview of the privacy laws is provided in later chapters.

Although most security experts would put regulatory compliance as the primary driver for data anonymization, this has been listed as the last driver in this book as the increasing risk of misuse of personal data by insiders and increasing operational risks adopted by businesses led governments and regulatory bodies to pass data privacy legislation.

Will Procuring and Implementing a Data Anonymization Tool by Itself Ensure Protection of Privacy of Sensitive Data?

From a data privacy protection perspective, data anonymization is only one of the popular approaches used. Other approaches like data loss prevention, data tokenization, etc., may also be used for specific data privacy protection requirements.

Data anonymization addresses data privacy protection by hiding the personal dimension of data or information. However, the implementation of only data anonymization (using data anonymization tools) without the support of policies, processes, and people will be inadequate.

There are companies who have used SQL scripts efficiently to encrypt data and have seen a fairly successful data anonymization implementation, although on a smaller scale. There are also companies whose initiatives have failed after procurement of the best data masking tool on the market. For protection from misuse of personal data, processes and policies need to come together along with the anonymization tool and the human aspect.

Employee training and increasing the awareness of information security and privacy guidelines and policies have played a positive role in enterprises being able to bring down insider data breaches due to negligence.

Some of the reasons for limited success or failure of data anonymization implementation include the following:

Ambiguity of Operational Aspects

Important decisions such as who is supposed to mask data, who can see unmasked data, and who can use the masking tool are not clearly defined.

Allowing the Same Users to Access Both Masked and Unmasked Environments

There are organizations that allow developers/testers/contractors access to unmasked data and have the same personnel mask these data and further use the masked or anonymized data only for development/testing on their premises. This defeats the purpose of anonymization.

Lack of Buy-In from IT Application Developers, Testers, and End-Users

Many implementations do not use a practical approach to data anonymization and do not secure the necessary buy-in from IT application developers, testers, and end-users before implementation and

as a result end up with testers who refuse to test with masked data as they are not realistic. A practical approach implies alignment with the organization's localized processes and procedures associated with its businesses, IT applications, and dataflow.

Compartmentalized Approach to Data Anonymization

Many large enterprises have different departments using different anonymization tools and approaches and the organization ends up not being able to perform an integration test with masked data.

Absence of Data Privacy Protection Policies or Weak Enforcement of Data Privacy Policies

Although most companies do know that customer names, date of birth, and national identification numbers need to be masked, there is no policy surrounding what type of data fields must be anonymized. Many companies lack the will to enforce the policy on employees not following the privacy policy guidelines, until a breach occurs. Without any supporting governance structure, data security and privacy policy, access control policies, and buy-in from a large section of IT employees that includes application developers and testers, data anonymization initiatives are bound to fail.

The next set of chapters provides a view on how data anonymization programs can be successfully implemented with supporting tools, processes, and people along with a set of patterns and antipatterns.

Benefits of Data Anonymization Implementation

Any security- or risk-related initiative will not result in an increase in generated revenues. It is only an insurance against "known" attacks that can bring down a business. Thus data anonymization implementation can help only in the protection of data privacy. It can ensure that nonproduction users cannot make use of the data while allowing them to continue using the application with same functionality as it exists in a production environment.

A piecemeal approach to data anonymization has its own pitfalls, however, data anonymization implemented in the right way with

all the supporting features across the enterprise has the following benefits:

- It reduces the likelihood of misuse of personal data by insiders and thereby the chance of litigation, especially when data are used in nonproduction environments.
- It increases adherence to data privacy laws and reduces hefty fines that may arise out of any misuse of personal data.
- More and more insurance companies are insuring their corporate customers only when they have a data security and privacy policy in place. Data anonymization, implemented the right way, should help reduce insurance premiums (by providing evidence of data security and privacy policy adherence) when insuring against data risks to the business.

A data anonymization program implemented at an enterprise level helps in standardization of data anonymization and privacy protection processes across the enterprise as well as reduction of operational cost of data anonymization.

Conclusion

Increasing incidences of insider data thefts and misuse of personal information of customers and employees have resulted in introduction of data privacy legislation by various governments and regulatory bodies. These pieces of legislation have made the cost of noncompliance and breach of personal data very expensive for businesses.

Although external attacks and hacker attacks on an enterprise can be prevented by network and physical security mechanisms, prevention of misuse of sensitive data can be achieved only by concerted data anonymization programs encompassing governance, processes, training, tools, techniques, and data security and privacy policy formulations.

References

1. Camouflage (http://doc.wowgao.com/ef/presentations/PPCamouflage.ppt)
2. *Guardian* (http://www.guardian.co.uk/healthcare-network/2011/may/04/personal-data-breaches-london-nhs-trusts-data)

3. *Guardian* (http://www.guardian.co.uk/healthcare-network/2011/may/04/biggest-threat-nhs-data-security-staff)
4. Datalossdb (http://www.datalossdb.org)
5. NIST (*Guide to Protecting the Confidentiality of Personally Identifiable Information*)
6. *USA Today* (cybercrime forum) (http://www.usatoday.com/tech/news/computersecurity/infotheft/2006-10-11-cybercrime-hacker-forums_x.htm)
7. *2010 Annual Study: U.S. Cost of a Data Breach* (Research conducted by Ponemon Institute. LLC and Sponsored by Symantec)