The Complete Book of Data Anonymization

From Planning to Implementation

Balaji Raghunathan

Infosys® Press

CRC Press
Taylor & Francis Group
AN AUERBACH BOOK

The Complete Book of Data Anonymization

From Planning to Implementation

Infosys® Press

In an initiative to promote authorship across the globe, Infosys Press and CRC Press have entered into a collaboration to develop titles on leading edge topics in IT.

Infosys Press seeks to develop and publish a series of pragmatic books on software engineering and information technologies, both current and emerging. Leveraging Infosys' extensive global experience helping clients to implement those technologies successfully, each book contains critical lessons learned and shows how to apply them in a real-world, enterprise setting. This open-ended and broad-ranging series aims to brings readers practical insight, specific guidance, and unique, informative examples not readily available elsewhere.

PUBLISHED IN THE SERIES

The Complete Book of Data Anonymization: From Planning to Implementation
Balaji Raghunathan

.NET 4 for Enterprise Architects and Developers
Sudhanshu Hate and Suchi Paharia

Process-Centric Architecture for Enterprise Software Systems
Parameswaran Seshan

Process-Driven SOA: Patterns for Aligning Business and IT
Carsten Hentrich and Uwe Zdun

Web-Based and Traditional Outsourcing
Vivek Sharma and Varun Sharma

IN PREPARATION FOR THE SERIES

Applying Resource Oriented Architecture: Using ROA to Build RESTful Web Services
G. Lakshmanan, S. V. Subrahmanya, S. Sangeetha, and Kumar M. Pradeep

Scrum Software Development
Jagdish Bhandarkar and J. Srinivas

Software Vulnerabilities Exposed
Sanjay Rawat, Ashutosh Saxena, and Ponnapalli K. B. Hari Gopal

The Complete Book of Data Anonymization

From Planning to Implementation

Balaji Raghunathan

Infosys® Press

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
AN AUERBACH BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20121205

International Standard Book Number-13: 978-1-4398-7731-9 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Introduction

As a data anonymization and data privacy protection solution architect, I have spent a good amount of time understanding how data anonymization, as a data privacy protection measure, is being approached by enterprises across different industrial sectors. Most of these enterprises approached enterprise-wide data anonymization more as an art than as a science.

Despite the initiation of data privacy protection measures like enterprise-wide data anonymization, a large number of enterprises still ran the risk of misuse of sensitive data by mischievous insiders. Though these enterprises procured advanced tools for data anonymization, many applications across the enterprise still used copies of actual production data for software development life cycle activities. The reasons for the less-than-expected success of data anonymization initiatives arose due to challenges arising from multiple quarters, ranging from technology to data to process to people.

This book intends to demystify data anonymization, identify the typical challenges faced by enterprises when they embark on enterprisewide data anonymization initiatives, and outline the best practices to address these challenges. This book recognizes that the challenges faced by the data anonymization program sponsor/manager are different from those of a data anonymization practitioner. The program sponsor's worries are more about getting the program executed on time

and on budget and ensuring the continuing success of the program as a whole whereas the practitioner's challenges are more technological or application-specific in nature.

Part I of this book is for the anonymization program sponsor, who can be the CIO or the IT director of the organization. In this part, this book describes the need for data anonymization, what data anonymization is, when to go in for data anonymization, how a data anonymization program should be scoped, what the challenges are when planning for this initiative at an enterprise-level scope, who in the organization needs to be involved in the program, which are the processes that need to be set up, and what operational aspects to watch out for.

Part II of this book is for the data anonymization practitioner, who can be a data architect, a technical lead, or an application architect. In this part, this book describes the different solution patterns and techniques available for data anonymization, how to select a pattern and a technique, the step-by-step approach toward data anonymization for an application, the challenges encountered, and the best practices involved.

This book is not intended to help design and develop data anonymization algorithms or techniques or build data anonymization tools. This book should be thought of more as a reference guide for data anonymization implementation.

Acknowledgments

More than an individual effort, this book is the result of the contributions of many people.

I would like to thank the key contributors:

Jophy Joy, from Infosys, for granting me permission to use all of the cartoons in this book. Jophy, who describes himself as a passionate "virus" for cartooning, has brought to life through his cartoons the lighter aspects of data anonymization, and has made the book more colorful.

Sandeep Karamongikar, from Infosys, for being instrumental in introducing me to the world of data anonymization, providing early feedback on the book, and ensuring executive support and guidance in publishing the book.

Venugopal Subbarao, from Infosys, for agreeing to review the book despite his hectic schedule, and providing expert guidance and comments, which helped shape this book.

Swaminathan Natarajan and Ramakrishna G. Reddy, from Infosys, for review of the book from a technical perspective.

Dr. Ramkumar Ramaswamy, from Performance Engineering Associates, as well as Ravindranath P. Hirolikar, Vishal Saxena, Shanmugavel S. and Santhosh G. Ramakrishna, from Infosys, for reviewing select chapters and providing their valuable comments.

Prasad Joshi, from Infosys, for providing executive support and guidance and ensuring that my official work assignments did not infringe on the time reserved for completing the book.

Dr. Pramod Varma, from Unique Identification Authority of India, for reading through the book and providing his valuable inputs on data privacy, and helping me with ideas for another book!!

Subu Goparaju and Dr. Anindya Sircar, from Infosys, for their executive guidance and support in publishing the book.

Sudhanshu Hate, from Infosys, and Parameshwaran Seshan, an independent trainer and consultant, for guiding me through the procedural aspects of getting the book published.

Dr. Praveen Bhasa Malla, from Infosys, for assisting me in getting this book published, right from the conceptual stage of the book.

Subramanya S.V., Dr. Sarma K.V.R.S., and Chidananda B. Gurumallappa, from Infosys, for their guidance in referencing external content in the book.

This book would not have been possible without the help received from Rich O'Hanley, Laurie Schlags, Michele A. Dimont, Deepa Jagdish, Kary A. Budyk, Elise Weinger, and Bill Pacheco, from Taylor & Francis. They patiently answered several of my queries and guided me through the entire journey of getting this book published.

I would also like to express my gratitude to Dr. Ten H. Lai, of Ohio State University, Cassie Stevenson, from Symantec, Susan Jayson, from Ponemon Institute, as well as Helen Wilson, from *The Guardian*, for providing me permission to reference content in my book.

I would like to dedicate this effort of writing a book to my father, P.K. Raghunathan, mother, Kalyani, wife, Vedavalli T.V., 8-year-old daughter, Samhitha, and 3-year-old son, Sankarshan, who waited for me for several weekends over a period of more than a year to finish writing this book and spend time with them. Their understanding and patience helped me concentrate on the book and get it out in due time.

Concerted efforts have been made to avoid any copyright violations. Wherever needed, permission has been sought from copyright owners. Adequate care has been taken in citing the right sources and references. However, should there be any errors or omissions, they are

inadvertent and I apologize for the same. I would be grateful for such errors to be brought to my attention so that they can be incorporated in the future reprints or editions of this work.

I acknowledge the proprietary rights of the trademarks and the product names of the companies mentioned in the book.

About the Author

Balaji Raghunathan has more than 15 years of experience in the software industry and has spent a large part of his working career in software architecture and information management. He has been with Infosys for the last 10 years.

In 2009, Raghunathan was introduced to data anonymization and ever since has been fascinated by this art and science of leaving users in doubt as to whether the data are real or anonymized. He is convinced that this is a valuable trick enterprises need to adopt in order to prevent misuse of personal data they handle and he has helped some of Infosys clients play these tricks systematically.

He is a TOGAF 8.0 and ICMG-WWISA Certified Software Architect and has worked on data anonymization solutions for close to two years in multiple roles. Prior to 2009, Raghunathan has been involved in architecting software solutions for the energy, utilities, publishing, transportation, retail, and banking industries.

Raghunathan has a postgraduate diploma in business administration (finance) from Symbiosis Institute (SCDL), Pune, India and has an engineering degree (electrical and electronics) from Bangalore University, India.

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