Bridging the Gap Between Privacy and Data Insight

Ulf Mattsson
CTO, Protegrity

ulf.mattsson[at]protegrity.com
Bridging the Gap Between Privacy and Data Insight
Ulf Mattsson, CTO Protegrity

- 20 years with IBM Research & Development & Global Services
- Started Protegrity in 1994 (Data Security)
- Inventor of 20+ patents
  - Encryption, Tokenization & Intrusion Prevention
- Member of
  - PCI Security Standards Council (PCI SSC)
  - American National Standards Institute (ANSI) X9
    - Encryption & Tokenization
  - International Federation for Information Processing
    - IFIP WG 11.3 Data and Application Security
  - NYOUG, ISACA, ISSA and Cloud Security Alliance (CSA)
Agenda

- HIPAA, PCI DSS & Privacy Laws
- Oracle’s Big Data Platform
- Big Data and Threats
  - Ways to Hack Big Data
- What’s the Problem with Securing Big Data?
  - Hadoop Beyond Kerberos
  - New Protection Techniques
  - Speed of Different Protection Methods
- Risk Adjusted Data Protection
  - De-Identifying Sensitive Data
  - Research Studies
- A Data Protection Methodology
  - Best Practices for Protecting Big Data
Balancing security and data insight

- **Big Data**
  - Sales & Marketing
  - Customer Profiles
  - Security Analysis
  - Business Improvement
  - Social Media

- **Regulations & Breaches**
- **Increased profits**
- **STOP**

Increased profits
Balancing security and data insight

- Tug of war between security and data insight
- Big Data is designed for access, not security
- Privacy regulations require de-identification which creates problems with privileged users in an access control security model
- Only way to truly protect data is to provide data-level protection
- Traditional means of security don’t offer granular protection that allows for seamless data use
Five Point Data Protection Methodology

Classify

*Determine what data is sensitive to your organization.*
1. Classify

Data is classified as sensitive and must be protected in response to Laws and regulations such as PCI DSS, HIPAA, State Privacy Laws and others.

Companies may also have Intellectual Property and other Company Secrets that must be secured against the competition.

All companies have sensitive data about their Customers and about their Employees.
1. Classify: Examples of Sensitive Data

<table>
<thead>
<tr>
<th>Sensitive Information</th>
<th>Compliance Regulation / Laws</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit Card Numbers</td>
<td>PCI DSS</td>
</tr>
<tr>
<td>Names</td>
<td>HIPAA, State Privacy Laws</td>
</tr>
<tr>
<td>Address</td>
<td>HIPAA, State Privacy Laws</td>
</tr>
<tr>
<td>Dates</td>
<td>HIPAA, State Privacy Laws</td>
</tr>
<tr>
<td>Phone Numbers</td>
<td>HIPAA, State Privacy Laws</td>
</tr>
<tr>
<td>Personal ID Numbers</td>
<td>HIPAA, State Privacy Laws</td>
</tr>
<tr>
<td>Personally owned property numbers</td>
<td>HIPAA, State Privacy Laws</td>
</tr>
<tr>
<td>Personal Characteristics</td>
<td>HIPAA, State Privacy Laws</td>
</tr>
<tr>
<td>Asset Information</td>
<td>HIPAA, State Privacy Laws</td>
</tr>
</tbody>
</table>

Breach notification laws are often the catalyst for a deep audit in companies resulting in large fines.
HIPAA PHI: List of 18 Identifiers

1. Names
2. All geographical subdivisions smaller than a State
3. All elements of dates (except year) related to individual
4. Phone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including fingerprints
17. Full face photographic images
18. Any other unique identifying number
## PCI DSS (Payment Card Industry Data Security Standard)

| Build and maintain a secure network. | 1. Install and maintain a firewall configuration to protect data  
2. Do not use vendor-supplied defaults for system passwords and other security parameters |
|--------------------------------------|--------------------------------------------------------------------------------------------------|
| Protect cardholder data.             | 3. Protect stored data  
4. Encrypt transmission of cardholder data and sensitive information across public networks |
| Maintain a vulnerability management program. | 5. Use and regularly update anti-virus software  
6. Develop and maintain secure systems and applications |
| Implement strong access control measures. | 7. Restrict access to data by business need-to-know  
8. Assign a unique ID to each person with computer access  
9. Restrict physical access to cardholder data |
| Regularly monitor and test networks. | 10. Track and monitor all access to network resources and cardholder data  
11. Regularly test security systems and processes |
| Maintain an information security policy. | 12. Maintain a policy that addresses information security |
Discovery

Discover where the sensitive data is located, how it flows, who can access it, the performance requirements and other requirements for protection.
2. Discovery

The Discovery process peers into the enterprise to find sensitive data in preparation for delivering an optimal protection solution.

- Existing Sensitive Data
- New Sensitive Data
- Archived Data
2. Discovery in a large enterprise with many systems

Focus on systems that contain sensitive data
2. Discovery: Determine the context to the Business

- Retail
- Corporate IP
- Healthcare
- Corporate Firewall

System
System
System
System
System
System
System
2. Discover: Context to the Business and to Security

- Stores and e-commerce collecting transactions
- File Server containing IP
- Applications Processing Retail Transactions
- File Server landing zone for store and e-commerce transactions
- Sensitive data in Hadoop
- Sensitive data in columns in databases
- Corporate Firewall
- Research Databases

Customer Data Protection Solution Requirements
Protect

Protect the sensitive data at rest and in transit.
Big Data and The Insider Threat

Google fired engineer for privacy breach

David Barksdale, a Google engineer, was sacked earlier this year for improperly accessing the accounts of several Google users, Google confirms.

by Tom Krazit | September 14, 2010 5:27 PM PDT

Google confirmed on Tuesday that it fired an employee earlier this year for violating its policies on accessing the accounts of its users.

Earlier in the day, Gawker reported that David Barksdale, an engineer in Google's Seattle offices, used his position as a key engineer evaluating the health of Google's services to break into the Gmail and Google Voice accounts of several children. After parents of the children complained to Google, Gawker said Barksdale--who was not accused of anything with sexual overtones--was dismissed, and Google confirmed that move late Tuesday.

"We dismissed David Barksdale for breaking Google's strict internal privacy policies. We carefully control the number of employees who have access to our systems, and we regularly
Big Data and The Insider Threat

Google fired engineer for privacy breach

David Barksdale, an engineer in Google's Search team, was fired this year for violating its privacy policies.

Earlier in the day, Gawker reported that Barksdale was dismissed, and Google confirmed that he was.

"We dismissed David Barksdale earlier this year for violating its privacy policies. We carefully control the number of people who can access this data," Google spokesperson Matt Kamen said in a statement.

Barksdale improperly accessed information about Gmail and Google Voice users, and had access to the children's personal information.

Barksdale worked for Google for just over five years. He joined the company in 2010 and was based in Mountain View, California.
Privacy Laws (See List in Appendix)

- 54 International Privacy Laws
- 30 United States Privacy Laws, including
  - Financial Services - Gramm-Leach-Bliley Act (GLBA), Sarbanes-Oxley Act (SARBOX), USA PATRIOT ACT, PCI Data Security Standard, and the Basel II Accord (EU)
  - Healthcare and Pharmaceuticals - HIPAA (Health Insurance Portability and Accountability Act of 1996) and FDA 21 CFR Part 11
  - Infrastructure and Energy - Guidelines for FERC and NERC Cybersecurity Standards, the Chemical Sector Cyber Security Program and Customs-Trade Partnership Against Terrorism (C-TPAT)
  - Federal Government - Compliance with FISMA and related NSA Guidelines and NIST Standards
Woman gets Prison Time in Identity Theft
Oracle’s Big Data Platform

Software on Oracle Big Data Appliance

## Software Layout - Oracle Big Data Appliance

<table>
<thead>
<tr>
<th></th>
<th>Node 1</th>
<th>Node 2</th>
<th>Node 3</th>
<th>Node 4-18</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Master</strong></td>
<td>NameNode</td>
<td>Secondary NameNode</td>
<td>JobTracker</td>
<td>DataNode</td>
</tr>
<tr>
<td></td>
<td>Balancer</td>
<td>Cloudera Manager</td>
<td>MySQL Master</td>
<td>TaskTracker</td>
</tr>
<tr>
<td></td>
<td>HBase Master</td>
<td>Zookeeper</td>
<td>ODI Agent</td>
<td>HBase Region Server</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NoSQL DB Administration*</td>
<td>NoSQL DB Storage Node*</td>
</tr>
<tr>
<td><strong>Slave</strong></td>
<td>DataNode</td>
<td>Data Node</td>
<td>Data Node</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NoSQL DB Storage Node*</td>
<td>NoSQL DB Storage Node*</td>
<td>NoSQL DB Storage Node*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MySQL DB Slave</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Legend:**
- **Yellow**: Hadoop Processes
- **Purple**: MySQL
- **Teal**: HBase
- **Green**: Oracle NoSQL DB

Source: 04_Oracle_Big_Data_Appliance_Deep_Dive.pdf
Many Ways to Hack Big Data

BI Applications

Data Access Framework
- Pig
- Hive
- Sqoop
- Avro

Data Processing Framework (MapReduce)

Data Storage Framework (HDFS)

Hackers

Unvetted Applications
Or
Ad Hoc Processes

Privileged Users

Many Ways to Hack Big Data
Hadoop - Protection Beyond Kerberos

**Data Storage Framework (HDFS)**

- Pig
- Hive
- Sqoop
- Avro

**Data Processing Framework (MapReduce)**

- **Field level data protection** with **Policy** based access control and Monitoring

**Data Access Framework**

- **Field level data protection** with **Policy** based access control and Monitoring

**BI Applications**

- **Field level data protection** with **Policy** based access control and Monitoring; existing and new data

- **Volume Encryption** with **Policy** based access control of Files and Monitoring
What’s the Problem with Securing Big Data?

• Analytics
• Inter-node data movement
• Encryption
  • Data size
  • Data type
  • Performance (SLA)
  • Table scans
Render PAN* unreadable anywhere it is stored by using any of the following approaches:

- **Index tokens** and pads
- Strong cryptography with associated key-management processes and procedures
- One-way hashes based on strong cryptography
- Truncation

* : Primary Account Number (credit card number)
Reduction of Pain with New Protection Techniques

Pain & TCO

High

Strong Encryption
AES, 3DES

!@#$%a^.,mhu7///&*B(_+!@

Format Preserving Encryption
DTP, FPE

Input Value: 3872 3789 1620 3675

Format Preserving
Greatly reduced Key Management

Low

Vault-based Tokenization
8278 2789 2990 2789

No Vault

Vaultless Tokenization
8278 2789 2990 2789
Speed of Different Protection Methods

Transactions per second*

<table>
<thead>
<tr>
<th>Method</th>
<th>Speed (Transactions per second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vault-based Data Tokenization</td>
<td>I</td>
</tr>
<tr>
<td>Format Preserving Encryption</td>
<td>I</td>
</tr>
<tr>
<td>AES CBC Encryption Standard</td>
<td>I</td>
</tr>
<tr>
<td>Vaultless Data Tokenization</td>
<td>I</td>
</tr>
</tbody>
</table>

*: Speed will depend on the configuration
Protection Granularity: Field Protection

Encryption
- Reversible
- Policy Control (authorized / Unauthorized Access)
- Lacks Integration Transparency
- Complex Key Management
- Example: !@#$%a^.,mhu7///&*B(_+!@

Masking
- Not reversible
- No Policy, Everyone can access the data
- Integrates Transparently
- No Complex Key Management
- Example: 0389 3778 3652 0038
Protection Granularity: Field Protection

**Encryption**
- Reversible
- Policy Control (authorized / Unauthorized Access)
- Lacks Integration Transparency
- Complex Key Management
- Example: !@#$%^,.mhu7///&*B(-+!@

**Vaultless Tokenization / Pseudonymization**
- Reversible
- Policy Control (Authorized / Unauthorized Access)
- Integrates Transparently
- No Complex Key Management
- Business Intelligence Credit Card: 0389 3778 3652 0038

**Masking**
- Not reversible
- No Policy, Everyone can access the data
- Integrates Transparently
- No Complex Key Management
- Example: 0389 3778 3652 0038
“Tokenization Gets Traction”

Aberdeen has seen a steady increase in enterprise use of tokenization for protecting sensitive data over encryption.

Nearly half of the respondents (47%) are currently using tokenization for something other than cardholder data.

Over the last 12 months, tokenization users had 50% fewer security-related incidents than tokenization non-users.
New Data Security = More Creativity

![Graph showing the relationship between risk and access to data, with traditional access control on the low end and new data security practices on the high end. The graph indicates that new data security practices allow for more creativity at the edge.](image)
## De-Identified Sensitive Data

<table>
<thead>
<tr>
<th>Field</th>
<th>Real Data</th>
<th>Tokenized / Pseudonymized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Joe Smith</td>
<td>csu wusoj</td>
</tr>
<tr>
<td>Address</td>
<td>100 Main Street, Pleasantville, CA</td>
<td>476 srtcoetse, cysieondusbak, CA</td>
</tr>
<tr>
<td>Date of Birth</td>
<td>12/25/1966</td>
<td>01/02/1966</td>
</tr>
<tr>
<td>Telephone</td>
<td>760-278-3389</td>
<td>760-389-2289</td>
</tr>
<tr>
<td>E-Mail Address</td>
<td><a href="mailto:joe.smith@surferdude.org">joe.smith@surferdude.org</a></td>
<td><a href="mailto:eoe.nwuer@beusorpdqo.org">eoe.nwuer@beusorpdqo.org</a></td>
</tr>
<tr>
<td>SSN</td>
<td>076-39-2778</td>
<td>076-28-3390</td>
</tr>
<tr>
<td>CC Number</td>
<td>3678 2289 3907 3378</td>
<td>3846 2290 3371 3378</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>![Fingerprint Image]</td>
<td>Encrypted</td>
</tr>
<tr>
<td>Photo</td>
<td>![Photo Image]</td>
<td>Encrypted</td>
</tr>
<tr>
<td>X-Ray</td>
<td>![X-Ray Image]</td>
<td>Encrypted</td>
</tr>
<tr>
<td>Healthcare Data – Primary Care Data</td>
<td>Dr. visits, prescriptions, hospital stays and discharges, clinical, billing, etc.</td>
<td>Protection methods can be equally applied to the actual healthcare data, but not needed with de-identification</td>
</tr>
</tbody>
</table>
## Flexibility in Token Format Controls

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Input</th>
<th>Token</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit Card</td>
<td>3872 3789 1620 3675</td>
<td>8278 2789 2990 2789</td>
<td>Numeric</td>
</tr>
<tr>
<td>Credit Card</td>
<td>3872 3789 1620 3675</td>
<td>3872 3789 2990 3675</td>
<td>Numeric, First 6, Last 4 digits exposed</td>
</tr>
<tr>
<td>Credit Card</td>
<td>3872 3789 1620 3675</td>
<td>3872 qN4e 5yPx 3675</td>
<td>Alpha-Numeric, Digits exposed</td>
</tr>
<tr>
<td>Account Num</td>
<td>29M2009ID</td>
<td>497HF390D</td>
<td>Alpha-Numeric</td>
</tr>
<tr>
<td>Date</td>
<td>10/30/1955</td>
<td>12/25/2034</td>
<td>Date - multiple date formats</td>
</tr>
<tr>
<td>E-mail Address</td>
<td><a href="mailto:yuri.gagarin@protegrity.com">yuri.gagarin@protegrity.com</a></td>
<td><a href="mailto:empo.snaugs@svtiensnni.snk">empo.snaugs@svtiensnni.snk</a></td>
<td>Alpha Numeric, delimiters in input preserved</td>
</tr>
<tr>
<td>SSN</td>
<td>075672278 or 075-67-2278</td>
<td>287382567 or 287-38-2567</td>
<td>Numeric, delimiters in input</td>
</tr>
<tr>
<td>Binary</td>
<td>0x010203</td>
<td>0x123296910112</td>
<td>Non length preserving</td>
</tr>
<tr>
<td>Decimal</td>
<td>123.45</td>
<td>9842.56</td>
<td>Non length preserving</td>
</tr>
<tr>
<td>Alphanumeric Indicator</td>
<td>5105 1051 0510 5100</td>
<td>8278 2789 299A 2781</td>
<td>Position to place alpha is configurable</td>
</tr>
<tr>
<td>Invalid Luhn</td>
<td>5105 1051 0510 5100</td>
<td>8278 2789 2990 2782</td>
<td>Luhn check will fail</td>
</tr>
<tr>
<td>Multi-Merchant or</td>
<td>3872 3789 1620 3675</td>
<td>ID 1: 8278 2789 2990 2789</td>
<td>This supports delivery of a different token</td>
</tr>
<tr>
<td>Multi-Client ID</td>
<td></td>
<td>ID 2: 9302 8999 2662 6345</td>
<td>to different merchant or clients based on the</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>same credit card number.</td>
</tr>
</tbody>
</table>
Volume Encryption

Entire file is in the clear when analyzed

Protected with Volume Encryption
Volume Encryption + Gateway Field Protection

Granular Field Level Protection

Data Protection File Gateway

MapReduce

HDFS

Kerberos Access Control

Protected with Volume Encryption
Volume Encryption + Internal MapReduce Field Protection

Granular Field Level Protection

Hadoop Staging

MapReduce

HDFS

Kerberos Access Control

Protected with Volume Encryption

Granular Field Level Protection
Enforce

*Policies are used to enforce rules about how sensitive data should be treated in the enterprise.*
4. Enforce

The goal of policy enforcement is to;

1. Hide sensitive data from un-authorized users but disclose sensitive data to authorized users.
2. Deliver the minimum information to an individual or a process who needs the information to accomplish a task. Least Privilege by NIST.
3. Collect information about who is attempting to access sensitive data – both authorized and unauthorized.
## A Data Security Policy

<table>
<thead>
<tr>
<th><strong>What</strong></th>
<th><strong>What</strong> is the sensitive data that needs to be protected. <strong>Data Element.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>How</strong></td>
<td><strong>How</strong> you want to protect and present sensitive data. There are several methods for protecting sensitive data. Encryption, tokenization, monitoring, etc.</td>
</tr>
<tr>
<td><strong>Who</strong></td>
<td><strong>Who</strong> should have access to sensitive data and who should not. Security access control. <strong>Roles &amp; Members.</strong></td>
</tr>
<tr>
<td><strong>When</strong></td>
<td><strong>When</strong> should sensitive data access be granted to those who have access. Day of week, time of day.</td>
</tr>
<tr>
<td><strong>Where</strong></td>
<td><strong>Where</strong> is the sensitive data stored? This will be where the policy is enforced. At the protector.</td>
</tr>
<tr>
<td><strong>Audit</strong></td>
<td><strong>Audit</strong> authorized or un-authorized access to sensitive data. Optional audit of protect/unprotect.</td>
</tr>
</tbody>
</table>
4. Enforce

Data Protection Policy

Access Control

Enterprise Data Warehouse

Big Data Clusters

Privileged Users

Business Application Users

External User & Processes

Bad Guys

Corporate Firewall
4. Enforce

Access Control

Data Protection Policy

Enterprise Data Warehouse

Big Data Clusters

Unauthorized

Authorized

Business Application Users

Privileged Users

Corporate Firewall

External User & Processes

Bad Guys

Authorized

Unauthorized
4. Enforce

Access Control

Data Protection Policy

Enterprise Data Warehouse

Big Data Clusters

Corporate Firewall

Privileged Users

Authorized

Unauthorized

Business Application Users

External User & Processes

Unauthorized

Authorized

Bad Guys

protegery
Volume Encryption + Field Protection + Policy Enforcement
4. Authorized User Example

Presentation to requestor
Name: Joe Smith
Address: 100 Main Street, Pleasantville, CA

Policy Enforcement

Authorized

Does the requestor have the authority to access the protected data?

Protection at rest
Name: csu wuso
Address: 476 sra coetse, cysieondusbak, CA
4. Un-Authorized User Example

Privileged User, DBA, System Administrators

Presentation to requestor
Name: csu wusoj
Address: 476 srita coetse, cysemjdusbak, CA

Policy Enforcement

Request

Response

Does the requestor have the authority to access the protected data?

Not Authorized

Protection at rest
Name: csu wusoj
Address: 476 srita coetse, cysemjdusbak, CA
Monitor

A critically important part of a security solution is the ongoing monitoring of any activity on sensitive data.
Policy enforcement collects information in the form of audit logs about any activity on sensitive data.

Monitoring enables security personnel to gain insights on what’s going on with your sensitive data.

It enables the understanding of what’s normal and what’s not normal activity on sensitive data.
Best Practices for Protecting Big Data

- **Start Early** – Don’t wait until you have terabytes of sensitive data in Hadoop before starting your Big Data protection program.

- **Granular protection** in addition to access control and volume protection.

- Future proof your protection. **Select the optimal protection** for today and for the future.

- **Enterprise coverage** to ensure nothing is left vulnerable.

- **Protection against insider threat** is more important today than ever before. Can only achieve this through granular data protection techniques.

- You can protect highly sensitive data while in a way that is mostly **transparent to the analysis process**.

- **Policy based protection** provides a shield to your sensitive data while recording all events on that data.
How Protegrity Can Help

1. We can help you **Classify** the sensitive data that needs to be secured in your enterprise.

2. We can help you **Discover** where the sensitive data sits in your environment and design the optimal security solution.

3. We can help you **Protect** your sensitive data with a flexible set of protectors and protection methods.

4. We can help you **Enforce** policies that will enable business functions while preventing sensitive data from getting in the wrong hands.

5. We can help you **Monitor** activity on sensitive data to gain insights on abnormal behaviors.
Introduction to Protegrity

Proven enterprise data security software and innovation leader

- Sole focus on the protection of data

Growth driven by risk management and compliance

- PCI (Payment Card Industry)
- PII (Personally Identifiable Information)
- PHI (Protected Health Information) – HIPAA
- State and Foreign Privacy Laws, Breach Notification Laws

Successful across many key industries
Please contact me for more information


Info@protegrity.com

www.protegrity.com