Privacy-Preserving Analytics and Secure Multiparty Computation

Ulf Mattsson
Chief Security Strategist
www.Protegrity.com

(ISC)2 East Bay Chapter
Ulf Mattsson

- **Chief Security Strategist**, Protegrity
- **Chief Technology Officer**, Protegrity, Atlantic BT, and Compliance Engineering
- **Head of Innovation**, TokenEx
- **IT Architect**, IBM

- Develops **Industry Standards**
- **Inventor** of more than 70 issued US Patents

**Products and Services:**
- Data Encryption, Tokenization, and Data Discovery
- Cloud Application Security Brokers (CASB) and Web Application Firewalls (WAF)
- Security Operation Center (SOC) and Managed Security Services (MSSP)
- Robotics and Applications

**Payment Card Industry (PCI) Security Standards Council (SSC):**
1. Tokenization Task Force
2. Encryption Task Force, Point to Point Encryption Task Force
3. Risk Assessment
4. eCommerce SIG
5. Cloud SIG, Virtualization SIG
6. Pre-Authorization SIG, Scoping SIG Working Group

**Cloud Security Alliance**

**Quantum Computing**

**Cybersecurity and Cryptographic Solutions**

**Tokenization Management and Security**

**Cloud Management and Security**

**Quantum Computing Risk Study**

**American National Standard for Financial Services**

**Data Security: On Premise or in the Cloud**

**Data Privacy: De-Identification Techniques**

**Privacy-Preserving Analytics and Secure Multi-Party Computation**
### Applications & Services being migrated to the Cloud

<table>
<thead>
<tr>
<th>Category</th>
<th>Currently migrating/deployed in the cloud</th>
<th>Planning to migrate in next 1-3 years</th>
<th>Planning to migrate in next 12 months</th>
<th>Have already migrated but moving/plan to move out of the cloud</th>
<th>Have already migrated but moving/plan to move to a different cloud model</th>
<th>Currently building from scratch/greenfield development</th>
<th>No plans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disaster Recovery / High Availability</td>
<td>40%</td>
<td>20%</td>
<td>14%</td>
<td>3%</td>
<td>4%</td>
<td>5%</td>
<td>15%</td>
</tr>
<tr>
<td>Storage / Archive / Backup / File Server</td>
<td>42%</td>
<td>17%</td>
<td>15%</td>
<td>3%</td>
<td>4%</td>
<td>6%</td>
<td>12%</td>
</tr>
<tr>
<td>CRM / ERP / HRMS / LOB applications</td>
<td>41%</td>
<td>17%</td>
<td>14%</td>
<td>3%</td>
<td>3%</td>
<td>4%</td>
<td>18%</td>
</tr>
<tr>
<td>IoT connectivity and management</td>
<td>23%</td>
<td>17%</td>
<td>13%</td>
<td>3%</td>
<td>2%</td>
<td>5%</td>
<td>37%</td>
</tr>
<tr>
<td>BI / Data warehouse (DW) / Data Analytics</td>
<td>30%</td>
<td>15%</td>
<td>15%</td>
<td>4%</td>
<td>4%</td>
<td>7%</td>
<td>25%</td>
</tr>
</tbody>
</table>
State Comprehensive-Privacy Law Comparison

The CCPA Effect

Task Force Substituted for Comprehensive Bill
Bill Died in Committee or Postponed
None

Statute/Bill in Legislative Process:
- Introduced
- In Committee
- Cross Chamber
- Cross Committee
- Passed
- Signed
The Rise of Legislation

[Map showing various data protection acts and regulations around the world, with highlights on GDPR, CCPA, PIPEDA, LGPD, POPI, PDPA, and their respective locations.]
GDPR under "Schrems II"

Legal safeguards:
- AWS Sarl guarantees in its contract with Doctolib, a French company, that it will challenge any general access request from a public authority.

Technical safeguards:
- Technically the data hosted by AWS Sarl is encrypted.
  - AWS Sarl, a Luxembourg registered company.
- The key is held by a trusted third party in France, not by AWS.

Other guarantees taken:
- No health data.
  - The data hosted relates only to the identification of individuals for the purpose of making appointments.
- Data is deleted after three months.

In scope for PCI DSS?

**System 0**

Encrypted Cardholder data (CHD)

**System 1**

Encryption process

**System 2**

Encryption keys

Encrypted Cardholder data (CHD)

**System 3**

Encryption keys

**System 4**

Encrypted Cardholder data (CHD)

---

The following is **NOT** in scope for PCI DSS:

**The following are each in scope for PCI DSS:**

1. Systems performing **encryption** and/or decryption of cardholder data, and systems performing **key management** functions

2. **Encrypted** cardholder data that is **not isolated** from the encryption and decryption and **key** management processes

3. **Encrypted** cardholder data that is present on a system or media that also contains the decryption **key**

4. **Encrypted** cardholder data that is present in the same environment as the decryption **key**

5. **Encrypted** cardholder data that is **accessible** to an entity that also has access to the decryption **key**

https://blog.pcisecuritystandards.org
HYOK (Hold Your Own Key) vs AWS Key Management

Client Control of Key Management

Client-side Key Management with AWS (BYOK)

HSM with AWS Key Management (BYOK)

AWS hosted Key Management

CASB (HYOK)

Agent (HYOK)

Client Control of Data

High

Low

High - Low
A Data Security Gateway Can Protect Sensitive Data in Cloud and On-premise
AWS Key Management

AWS hosted Key Management (AWS Keys)
AWS Key Management

Client-side Key Management with AWS (BYOK)
AWS Key Management

HSM with AWS
Key
Management
(BYOK)
SASE Is One of the Fastest Growing Markets

35.1% 5-Year CAGR

USD Billions

<table>
<thead>
<tr>
<th>Year</th>
<th>Network</th>
<th>Security</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>1.3</td>
<td>2.1</td>
<td>3.4</td>
</tr>
<tr>
<td>2021</td>
<td>2.3</td>
<td>2.9</td>
<td>5.2</td>
</tr>
<tr>
<td>2022</td>
<td>3.4</td>
<td>3.7</td>
<td>7.1</td>
</tr>
<tr>
<td>2023</td>
<td>4.7</td>
<td>4.8</td>
<td>9.5</td>
</tr>
<tr>
<td>2024</td>
<td>6.0</td>
<td>6.2</td>
<td>12.2</td>
</tr>
<tr>
<td>2025</td>
<td>7.5</td>
<td>7.6</td>
<td>15.1</td>
</tr>
</tbody>
</table>
Cloud Security Logical Architecture
Cloud Security Architecture
Big Data Protection with Granular Field Level Protection for Google Cloud
Use Case (Financial Services) - Compliance with Cross-Border and Other Privacy Restrictions

Central Security Manager (ESA)

On-premise or hosted

CENTRAL CONTROL (US) – LOCAL DATA MANAGEMENT

Local Data Security Gateways (DSG)
- 200 million users
- 160 countries

DOMAIN IN EMEA

DOMAIN IN APAC

PROTEGERITY

Copyright © Protegrity Corp.
Protection of data in AWS S3 with Separation of Duties

- Applications can use de-identified data or data in the clear based on policies
- Protection of data in AWS S3 before landing in a S3 bucket

Separation of Duties

- Encryption Key Management
- Policy Enforcement Point (PEP)
Consistency

- Most firms are quite familiar with their on-premises encryption and key management systems, so they often prefer to leverage the same tool and skills across multiple clouds.
- Firms often adopt a “best of breed” cloud approach.

Trust

- Some customers simply do not trust their vendors.

Vendor Lock-in and Migration

- A common concern is vendor lock-in, and an inability to migrate to another cloud service provider.
<table>
<thead>
<tr>
<th></th>
<th>Snowflake</th>
<th>Amazon Redshift</th>
<th>Azure Synapse Analytics</th>
<th>Google BigQuery</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sharing</strong></td>
<td>Easy between different accounts.</td>
<td>Multiple data output formats</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Flexibility</strong></td>
<td>Flexible</td>
<td>Less flexible</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Scale</strong></td>
<td>Instant scaling. Easy to maintain</td>
<td>Scale up and scaledown manually</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Management</strong></td>
<td>Easy to set up. Automated</td>
<td>Periodic vacuuming tables. Min</td>
<td>Automatic management.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>maintenance</td>
<td>administration</td>
<td>Intuitive</td>
<td></td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>Rows. Support for JSON</td>
<td>Column-oriented</td>
<td></td>
<td>columnar</td>
</tr>
<tr>
<td><strong>Roll back</strong></td>
<td></td>
<td>Roll-back on transactions</td>
<td></td>
<td>Cannot roll back on transactions</td>
</tr>
<tr>
<td><strong>Integration</strong></td>
<td>Ingesting fast. On-premise doesn’t</td>
<td>Largest cloud ecosystem. Much latency</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>integrate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Speed</strong></td>
<td></td>
<td>2 times faster than other</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Easy</strong></td>
<td></td>
<td>User-friendly. Datalakes, easy</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Price starts</strong></td>
<td>Storage costs separate. $2.01 per</td>
<td>Pay as you use model. $0.25 per hour</td>
<td>Not cost-effective</td>
<td>Complicated. Cost separate for</td>
</tr>
<tr>
<td></td>
<td>hour.</td>
<td></td>
<td>as others</td>
<td>storage. Queries cost $5/TB,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>add quickly</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td></td>
<td>$306 per TB per month</td>
<td></td>
<td>$20 per TB per month</td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td>Secure views and user-defined</td>
<td>Rich cloud services</td>
<td>B2B identity management</td>
<td></td>
</tr>
<tr>
<td></td>
<td>functions</td>
<td></td>
<td>with Oauth</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Encryption for client and server</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Column-level access control</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td></td>
<td>Needs improvement.</td>
<td>Great support</td>
<td></td>
</tr>
<tr>
<td><strong>Stability</strong></td>
<td></td>
<td>Robust. Some issues with stability</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Max columns</strong></td>
<td>1,600 columns in a single table</td>
<td></td>
<td></td>
<td>10,000 columns</td>
</tr>
<tr>
<td><strong>Workloads</strong></td>
<td></td>
<td>Analytical not transactional</td>
<td></td>
<td>Great at big chunks in a small</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>time. Data scientists and ML</td>
</tr>
</tbody>
</table>
Secure Multi-Party Computation
Secure Multi-Party Computation (MPC)

Private multi-party machine learning with MPC

Using MPC, different parties send encrypted messages to each other, and obtain the model $F(A,B,C)$ they wanted to compute without revealing their own private input, and without the need for a trusted central authority.
Case Study – HE and Securely sharing sensitive information

An example from the healthcare domain.

The recent ability to fully map the human genome has opened endless possibilities for advances in healthcare.

1. Data from DNA analysis can test for genetic abnormalities, empower disease-risk analysis, discover family history, and the presence of an Alzheimer’s allele.
   - But these studies require very large DNA sample sizes to detect accurate patterns.

2. However, sharing personal DNA data is a particularly problematic domain.
   - Many citizens hesitate to share such personal information with third-party providers, uncertain of if, how and to whom the information might be shared downstream.

3. Moreover, legal limitations designed to protect privacy restrict providers from sharing this data as well.

4. HE techniques enable citizens to share their genome data and retain key privacy concerns without the traditional all-or-nothing trust threshold with third-party providers.
Analytics and AI
Increased need for data analytics drives requirements.

Internal and Individual Third-Party Data Sharing

- Data Lake, ETL, Files...
- Data Pipeline
- Data Privacy

Secure Multi Party Computation

- Analytics, Data Science, AI and ML
- Data Pipeline
- Cloud
- On-premises
- Data Collaboration
- External Data
- Internal Data

Protected data fields
Policy Enforcement Point (PEP)
Encryption Key Management
Global Hadoop Big Data Analytics Market

Real-time data is significant in global datasphere

Between 2018 and 2025 the size of real-time data in the global datasphere is expected to expand tenfold, from five zettabytes to 51 zettabytes.

Source: Adapted from Maximize Market Research
Analytics and AI

Feelings about Impact of New Technologies

<table>
<thead>
<tr>
<th>Area</th>
<th>Positive</th>
<th>Negative</th>
<th>Don't know</th>
<th>Equal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personalized medicine</td>
<td>50</td>
<td>10</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>Driverless cars</td>
<td>44</td>
<td>20</td>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>AI</td>
<td>44</td>
<td>20</td>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>Gene editing</td>
<td>41</td>
<td>14</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td>Blockchain</td>
<td>35</td>
<td>18</td>
<td>25</td>
<td>21</td>
</tr>
</tbody>
</table>

Source: Adapted from Edelman Trust Barometer

What did we do before Machine Learning?

Simple Pattern Matching

Statistical Methods

Rules and First Order Logic (FoL)

Artificial Intelligence and Machine Learning

The Big Picture
- Artificial Intelligence
- Machine Learning
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Reinforcement Learning

Artificial Intelligence
Increased need for Data Analytics

Reduce Risk
• Secure AI & ML

Use-cases
• Analysis
• Insight
• Dashboarding
• Reporting
• Predictions
• Forecasts
• Simulation
• Optimization

Values
• Savings
• Revenue add

Anonymization to minimize the risk of identification

Examples in Banking Credit Card Approval,
• Reducing the risk from 26% down to 8%
• 98% accuracy compared to the Initial Model
Secure AI & ML

Use-cases

• Analysis
• Insight
• Dashboarding
• Reporting
• Predictions
• Forecasts
• Simulation
• Optimization

Values

• Savings
• Revenue add

Anonymization to minimize the risk of identification

• Examples in Banking Credit Card Approval,
• Reducing the risk from 26% down to 8%
• 98% accuracy compared to the Initial Model
Gartner Hype Cycle for Emerging Technologies, 2020

AI & ML
Algorithmic Trust Models Can Help Ensure Data Privacy

Emerging technologies tied to algorithmic trust include:

1. Secure access service edge (SASE)
2. Explainable AI
3. Responsible AI
4. Bring your own identity
5. Differential privacy
6. Authenticated provenance
<table>
<thead>
<tr>
<th>Product</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataiku</td>
<td>4.40</td>
</tr>
<tr>
<td>SAS</td>
<td>4.36</td>
</tr>
<tr>
<td>TIBCO Software</td>
<td>4.33</td>
</tr>
<tr>
<td>KNIME</td>
<td>4.32</td>
</tr>
<tr>
<td>IBM</td>
<td>4.27</td>
</tr>
<tr>
<td>RapidMiner</td>
<td>4.27</td>
</tr>
<tr>
<td>MathWorks</td>
<td>4.25</td>
</tr>
<tr>
<td>Alteryx</td>
<td>4.23</td>
</tr>
<tr>
<td>Databricks</td>
<td>4.13</td>
</tr>
<tr>
<td>DataRobot</td>
<td>4.12</td>
</tr>
<tr>
<td>Microsoft</td>
<td>4.04</td>
</tr>
<tr>
<td>Altair</td>
<td>3.98</td>
</tr>
<tr>
<td>Domino</td>
<td>3.97</td>
</tr>
<tr>
<td>H2O.ai</td>
<td>3.87</td>
</tr>
<tr>
<td>Google</td>
<td>3.86</td>
</tr>
</tbody>
</table>
Gartner MQ for Data Science and Machine Learning Platforms

Data and analytics pipeline, including all the following areas:

1. Data ingestion
2. Data preparation
3. Data exploration
4. Feature engineering
5. Model creation and training
6. Model testing
7. Deployment
8. Monitoring
9. Maintenance
10. Collaboration

Gartner MQ for Data Science and Machine Learning Platforms, 2021

Data and analytics pipeline, including all the following areas:

1. Data ingestion
2. Data preparation
3. Data exploration
4. Feature engineering
5. Model creation and training
6. Model testing
7. Deployment
8. Monitoring
9. Maintenance
10. Collaboration
Responsible AI

Aims to address the Trust and Ethics related to decisions made by AI models

- Ethics & Fairness
- Interpretability
- Privacy

Data Generalization
- Differential Privacy
- k-anonymity
- l-diversity
- t-closeness

Synthetic Data
- GANs
- VAE
- GANs with DP

Private AI Models
- Torch with DP
- Scikit-learn DP

Confidential AI

Aims to enable centralized and decentralized AI consortia safely

ML on Encrypted Data
- Queries as-a-service
- AI Models as-a-service

Federated Learning
- Training, testing, analytics—they're all tasks we can tackle privately and securely with federation
Use Case: Insilico Medicine

An alternative to animal testing for research and development programs in the pharmaceutical industry.

- By using **artificial intelligence** and deep-learning techniques, Insilico is able to analyze how a compound will affect cells and what drugs can be used to treat the cells in addition to possible side effects.

A comprehensive drug discovery engine, which utilizes **millions of samples and multiple data types** to discover signatures of disease and identify the most promising targets for billions of molecules that already exist or can be generated de novo with the desired set of parameters.
Machine Learning Model Lifecycle - Example

1. Define the model: using the Sequential or Model class and add the layers
2. Compile the model: call compile method and specify the loss, optimizer and metrics
3. Train the model: call fit method and use training data
4. Evaluate the model: call evaluate method and use testing data to evaluate trained model
5. Get predictions: use predict method on new data for predictions
Specify Access Control and Data Protection to Use

Review Use Cases and Types of Data

Implement

1. Dynamic Masking
2. Tokenization
3. Encryption
Data Protection Techniques
What Data Protection Technique do I need?

**Transparent DB Encryption**
Native database controls for encrypting data at rest

**Access Control**
Database views are native access control tools that limit the data that can be accessed

**Monitoring**
Transactional auditing and monitoring that provides greater context to who, what, and how data is being accessed

**Masking**
Presentation layer data protection that does not change the data at rest or in transit

**Examples of a few Privacy-Preserving Techniques**

**Tokenization**
Data deidentification that provides superior data protection

---

High Usability

High Security
Unlock the Potential of Data Security
- Data Security Governance Stakeholders

**CDO:** Data is our greatest monetization asset.

**CFO:** Let’s manage data as a **real** asset.

**CISO:** I don’t have enough resources to protect all this data.

**DPO:** And make sure we protect our customers’ privacy.

Source: Gartner
Protect data in ways that are transparent to business processes and compliant to regulations.

**CDO:** Data is our greatest monetization asset.

**CFO:** Let’s manage data as a real asset.

**CISO:** I don’t have enough resources to protect all this data.

**DPO:** And make sure we protect our customers’ privacy.

---

Source: Gartner
What are the Drivers for these People?

**Permissions**

- **Opportunities**
- **Risk Management**
- **Policies**
- **Balance**
- **Breaches**
- **Controls**

**Sources:**
- Adapated from Gartner

**CEOs**
- Focusing on **Short term Revenue**
- May not be at the same company next year and potentially not interested in longer projects

**CFOs**
- Let's manage data as a **real** asset
- IT Cost resonates with CIO

**CIOs**
- CIO is focusing on Short term Revenue at may allow Analytics on Unprotected Data
- CISO: I don't have enough resources to protect all this data.

**C&IOs**
- CDO: Data is our greatest monetization asset
- CMO is focusing on Short term Revenue at may allow Analytics on Unprotected Data
- Customer Trust and Liability resonates with DPO

**CMOs**
- CMO is focusing on Short term Revenue at may allow Analytics on Unprotected Data
- IT Cost and Liability resonates with CFO

**CFOs**
- Let's manage data as a **real** asset
- IT Cost resonates with CIO

**CISOs**
- CISO focus on Regulations and may not be at the same company next year and potentially not interested in longer projects

**DPOs**
- DPO: And make sure we protect our customers’ privacy

**Customer Trust**
- Customer Trust and Liability resonates with DPO

**Source:** Adapted from Gartner
What is the Cost of Implementing different Data Protection?

**Implementation Effort**

<table>
<thead>
<tr>
<th>What you need to do</th>
<th>Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create Data Element</td>
<td></td>
</tr>
<tr>
<td>Create Roles</td>
<td></td>
</tr>
<tr>
<td>Create Policy</td>
<td></td>
</tr>
<tr>
<td>Create Masks</td>
<td></td>
</tr>
<tr>
<td>Instrument DSG</td>
<td></td>
</tr>
<tr>
<td>Instrument Databases</td>
<td></td>
</tr>
<tr>
<td>Tokenize Data at Rest</td>
<td></td>
</tr>
</tbody>
</table>

Some Effort Required

No Effort Required

- Tokenization
- Database
- Gateway
- Mask
- Policy

Instrumentation

Instrumentation
What Data is Sensitive?

What Data is Regulated?

- EU GDPR
- US California CCPA / CPRA
- PCI DSS
- US HIPAA
Are my Use Cases involving Multi-party Computing and Homomorphic Encryption?

Do I need to plan for Quantum Computing?

Use Cases involving Analytics

Pseudonymization Of Identifiers

2-way algorithms

1-way algorithms

Anonymization Of Attributes

Data Format and Type Preserving

Differential Privacy (DP)

K-anonymity model

Static Derivation

Synthetic Data

HE in relation to QC and ML

Homomorphic Encryption (HE)

Public Key Encryption (PKE)

Lattice based encryption

Trusted Execution Environments (TEE)

Machine Learning (ML)

Analytics

Quantum Computing (QC)
Secure AI – Use Case with Synthetic Data

<table>
<thead>
<tr>
<th>Address</th>
<th>Name</th>
<th>Email</th>
<th>Phone</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Fully Synthetic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Artificially generated new data points

<table>
<thead>
<tr>
<th>Partially Synthetic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Artificially generated new data points

Derived new data points
Public Cloud (IaaS) Risk, Complexity & Cost

CIO — Security and Complexity Are Top Challenges

Top Challenges Using Multiple Public Cloud Infrastructure (IaaS) Providers
Up to Three Selections Allowed

- Increased security risks: 46%
- Increased complexity operating/ administer multiple techs: 45%
- Increased complexity of managing multiple bills: 35%
- More expensive: 33%
- Hard to find IT service providers with the skills for all my cloud properties: 31%
- Diversification/dilution of internal skills: 30%
- Higher need to use external skills: 30%
- Requires the use of 3rd-party cloud management tools: 28%
Homomorphic encryption (HE)

HE depicted in a client-server model

- The client sends encrypted data to a server, where a specific analysis is performed on the encrypted data, without decrypting that data.

- The encrypted result is then sent to the client, who can decrypt it to obtain the result of the analysis they wished to outsource.
## Use Cases for Secure Multi Party Computation & Homomorphic Encryption (HE)

### Business models and application domains:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Genomics</th>
<th>Health</th>
<th>National Security</th>
<th>Education</th>
<th>Social Security</th>
<th>Business Analytics</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Topics</td>
<td>GWAS</td>
<td>billing and reporting</td>
<td>smart grid</td>
<td>school dropouts</td>
<td>credit history</td>
<td>prediction</td>
<td>storage, sharing</td>
</tr>
<tr>
<td>Data Owner</td>
<td>medical institutions</td>
<td>clinics and hospitals</td>
<td>nodes and network</td>
<td>schools, welfare</td>
<td>government</td>
<td>business owners</td>
<td>clients</td>
</tr>
<tr>
<td>Why HE?</td>
<td>HIPAA</td>
<td>cyber insurance</td>
<td>privacy</td>
<td>FERPA</td>
<td>cyber crimes</td>
<td>data are valuable</td>
<td>untrusted server</td>
</tr>
<tr>
<td>Who pays?</td>
<td>health insurance</td>
<td>hospital</td>
<td>energy company</td>
<td>DoE</td>
<td>government</td>
<td>business owners</td>
<td>clients</td>
</tr>
</tbody>
</table>

http://homomorphiccryption.org
Use case – Retail - Data for Secondary Purposes

Large aggregator of credit card transaction data.

Open a new revenue stream

• Using its data with its business partners: retailers, banks and advertising companies.
• They could help their partners achieve better ad conversion rate, improved customer satisfaction, and more timely offerings.
• Needed to respect user privacy and specific regulations. In this specific case, they wanted to work with a retailer.
• Allow the retailer to gain insights while protecting user privacy, and the credit card organization’s IP.
• An analyst at each organization’s office first used the software to link the data without exchanging any of the underlying data.

Data used to train the machine learning and statistical models.

• A logistic and linear regression model was trained using secure multi-party computation (SMC).
• In the simplest form SMC splits a dataset into secret shares and enables you to train a model without needing to put together the pieces.
• The information that is communicated between the peers is encrypted at all times and cannot be reverse engineered.

With the augmented dataset, the retailer was able to get a better picture of its customers buying habits.
Use case - Financial services industry

Confidential financial datasets which are vital for gaining significant insights.

- The use of this data requires navigating a minefield of private client information as well as sharing data between independent financial institutions, to create a statistically significant dataset.
- Data privacy regulations such as CCPA, GDPR and other emerging regulations around the world
- Data residency controls as well as enable data sharing in a secure and private fashion.

Reduce and remove the legal, risk and compliance processes
- Collaboration across divisions, other organizations and across jurisdictions where data cannot be relocated or shared
- Generating privacy respectful datasets with higher analytical value for Data Science and Analytics applications.
Use case: Bank - Internal Data Usage by Other Units

A large bank wanted to broaden access to its **data lake** without compromising data **privacy**, preserving the data’s **analytical value**, and at reasonable infrastructure costs.

- Current approaches to **de-identify data** did **not** fulfill the **compliance** requirements and business needs, which had led to several bank **projects being stopped**.
- The issue with these techniques, like **masking**, **tokenization**, and **aggregation**, was that they did **not** sufficiently **protect** the data **without overly degrading data quality**.

This approach allows creating privacy protected datasets that retain their analytical value for Data Science and business applications.

A plug-in to the organization’s analytical pipeline to **enforce the compliance policies before the data was consumed by data science** and business teams from the data lake.

- The analytical quality of the data was preserved for machine learning purposes by-using AI and leveraging privacy models like **differential privacy and k-anonymity**.

Improved data access for teams **increased the business’ bottom line** without adding excessive infrastructure costs, while **reducing the risk** of-consumer information exposure.
Trusted execution environments

Trusted Execution Environments (TEEs) provide secure computation capability through a combination of special-purpose hardware in modern processors and software built to use those hardware features.

The special-purpose hardware provides a mechanism by which a process can run on a processor without its memory or execution state being visible to any other process on the processor,
• not even the operating system or other privileged code.

Computation in a TEE is not performed on data while it remains encrypted.
• Typically, the memory space of each TEE (enclave) application is protected from access
  • AES-encrypted when and if it is stored off-chip.

Usability is low and products/services are emerging in MS Azure, IBM’s cloud service Amazon AWS (late 2020)*

*: Source: http://publications.officialstatistics.org
In differential privacy, the concern is about privacy as the relative difference in the result whether a specific individual or entity is included in the input or excluded.
For k-anonymity to be achieved, there need to be at least k individuals in the dataset who share the set of attributes that might become identifying for each individual.

K-anonymity might be described as a ‘hiding in the crowd’ guarantee: if each individual is part of a larger group, then any of the records in this group could correspond to a single person.

This second table shows the data anonymised to achieve k-anonymity of k = 3, as you can see this was achieved by generalising some quasi-identifier attributes and redacting some others.
# Data protection techniques: Deployment on-premises, and clouds

<table>
<thead>
<tr>
<th>Privacy enhancing data de-identification terminology and classification of techniques</th>
<th>Data Warehouse</th>
<th>Centralized</th>
<th>Distributed</th>
<th>On-premises</th>
<th>Public Cloud</th>
<th>Private Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>De-identification techniques</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tokenization</td>
<td>Vault-based tokenization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vault-less tokenization</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cryptographic tools</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Format preserving encryption</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Homomorphic encryption</td>
<td></td>
<td></td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Suppression techniques</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Masking</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Hashing</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td><strong>Formal privacy measurement models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Differential Privacy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Server model</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Local model</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td><strong>K-anonymity model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-diversity</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>T-closeness</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
</tbody>
</table>
Clear Text Data Source

2-way Pseudonymization Of Identifiers

Format Preserving Encryption (FPE)

Tokenization

Format Preserving Encryption (FPE)

Algorithmic Random Noise added

Format Preserving Encryption (HE)

Computing on encrypted data

Hashing

Data store

Differential Privacy (DP)

Static Data Masking

K-anonymity model

Static Derivation

1-way Anonymization Of Attributes

Format Preserving

Very slow

Fastest Fast Slow

Fastest Fast Slow

Fastest Fast Slow

User

Dynamic Data Masking
Quantum Computers?

- **Quantum** computers and other strong computers can break algorithms and patterns in encrypted data.
- **We can instead use random** numbers to secure sensitive data.
- **Random** numbers are not based on an algorithm or pattern that computers can break.

Tech giants are building their own machines and speeding to make them available to the world as a cloud computing service. In the competition: IBM, Google, Microsoft, Intel, Amazon, IonQ, Quantum Circuits, Rigetti Computing
Lower Risk and Higher Productivity with More Access to More Data

Risk

High -
User Productivity

More Access to Data

High Risk (Clear Data)

Low Risk (Tokens)

Access to Data

High Risk Clear Data

Real Data
Joe Smith
100 Main Street, Pleasantville, CA
12/25/1966
760-278-3389
joe.smith@surfordude.org
076-30-2778
3079 2289 3907 3378
www.surfordude.com

Low Risk Tokens

Tokenized / Pseudonymized
csu woxoj
476 srta coeze, cysieondubak, CA
01/02/1966
760-389-2289
coe.nwuer@beusorospdo.org
076-28-3350
3846 2250 3371 3378
www.sheytntao.com

Encrypted

Encrypted

Encrypted

Protection methods can be equally applied to the actual data, but not needed with de-identification
**Non-reversible Data Transformations**

**Clear Text Data Source**

### Example of Data Generalization

<table>
<thead>
<tr>
<th>Source data:</th>
<th>Output data:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient</strong></td>
<td><strong>Age</strong></td>
</tr>
<tr>
<td>173965429</td>
<td>57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Generalization</strong></th>
<th><strong>Generalization</strong></th>
</tr>
</thead>
</table>

- **Hashing**
- **Anonymization Of Attributes**
- **K-anonymity model**
- **Format Preserving**
- **Static Data Masking**
- **Differential Privacy (DP)**
- **Noise added**
- **Synthetic Data**

**Data store**

**User**

**Dynamic Data Masking**
<table>
<thead>
<tr>
<th>Technique name</th>
<th>Use Case / User Story</th>
<th>Data protected in</th>
<th>Data truthfulness at record level</th>
<th>Applicable to types of attributes</th>
<th>Reduces the risk of</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Transit</td>
<td>Use</td>
<td>Storage</td>
<td></td>
</tr>
<tr>
<td>Pseudonymization</td>
<td>Tokenization</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cryptographic tools</td>
<td>Deterministic encryption</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Order-preserving encryption</td>
<td>Partially</td>
<td>Partially</td>
<td>Partially</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Homomorphic encryption</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Masking</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Suppression</td>
<td>Local suppression</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Record suppression</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Sampling</td>
<td>Partially</td>
<td>Partially</td>
<td>Partially</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Generalization</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Rounding</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Top/bottom coding</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Partially</td>
</tr>
<tr>
<td></td>
<td>Noise addition</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Randomization</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Micro aggregation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Privacy models</td>
<td>Differential privacy</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>K-anonymity</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Source: INTERNATIONAL STANDARD ISO/IEC 20889

Copyright © Protegrity Corp.
Difference between encryption and tokenization techniques

Source: INTERNATIONAL STANDARD ISO/IEC 20889
Examples of operational aspects of different tokenization techniques

Source: INTERNATIONAL STANDARD ISO/IEC 20889

- **Vault-less**
  - Small, static
  - No replication required
  - No collisions
  - Minimum impact on performance and scalability. All operations in memory.

- **Vault-based Dynamic**
  - Large, expanding
  - Replication required
  - Prone to collisions
  - Will impact performance and scalability

- **Vault-based Pre-generated**
  - Large, static
  - No replication required
  - No collisions
  - Will impact performance and scalability. Faster than dynamic approach.

- **Replication & Collisions**
What Data Protection Technique is Fastest?

Examples of Speed for Different Data Protection Techniques

- **Vault-based Data Tokenization**: Example of 1 k/s with a centralized Token Vault on Oracle

- **Format Preserving Encryption**: Example of one FPE encryption implements 10 rounds of AES

- **AES CBC Encryption Standard**: Example of Vaultless Tokenization performance is comparable to AES

- **Vaultless Data Tokenization**:
  - AWS lambda example: 180 million/s
  - Teradata example: 10 million/s
  - Linux on Intel® Xeon® Processor E5 Family example: 200 k/s

*: Speed will depend on the configuration
For k-anonymity to be achieved, there need to be at least k individuals in the dataset who share the set of attributes that might become identifying for each individual.

K-anonymity might be described as a ‘hiding in the crowd’ guarantee: if each individual is part of a larger group, then any of the records in this group could correspond to a single person.

This second table shows the data anonymised to achieve k-anonymity of k = 3, as you can see this was achieved by generalising some quasi-identifier attributes and redacting some others.

### Differential Privacy (DP)

- **Clear**: Primary DB → Protected Curator* Filter → Clear
- **Protected Curator* Filter**: Contains sensitive data
- **Cleanser Filter**: Removes sensitive data

### k-Anonymity Model

- **Clear**: DB → Cleanser Filter
- **Protected Curator* Filter**: Contains k-anonymised data
11 Published International Privacy Standards

ISO Privacy Standards

- **Techniques**
  - 20889 IS Privacy enhancing de-identification terminology and classification of techniques
  - 27701 IS Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management - Requirements and guidelines
  - 27018 IS Code of practice for protection of PII in public clouds acting as PII processors
  - 29000 IS Privacy framework
  - 29101 IS Privacy architecture framework
  - 29134 IS Guidelines for Privacy impact assessment
  - 29190 IS Privacy capability assessment model
  - 29191 IS Requirements for partially anonymous, partially unlinkable authentication

- **Management**
  - 29151 IS Code of Practice for PII Protection

- **Cloud**
  - 19608 TS Guidance for developing security and privacy functional requirements based on 15408

- **Framework**
  - 27550 TR Privacy engineering for system lifecycle processes

- **Impact**
  - 27000 TR: Technical Report

- **Requirements**
  - TS: Technical Specification

- **Process**
  - IS: International Standard

Guidelines to help comply with ethical standards
References A:


Thank You!

Ulf Mattsson
Chief Security Strategist
www.Protegrity.com
Was your organization prepared to enable a fully remote workforce before the pandemic?

- Yes: 56.02%
- No: 31.95%
- Unsure: 12.04%

Has your organization loosened its security policies and settings now that most people are presumably working from home?

- Yes, we have loosened things: 11.16%
- No, more people are working from home, but we haven't changed anything in our security: 49.45%
- No, we have tightened our security policy and settings: 34.57%
- In my organization, there has not been an increase in the amount of people working from home now: 4.81%
How to protect different types of data with encryption and tokenization

Use Case

Simple

Complex

Encryption of Files

Tokenization of Fields

Type of Data

Un-structured

Structured

Payment Card Information

Card Holder Data

Personal Information (PI*) or Personally Identifiable Information (PII)

Protected Health Information

PHI

Personally Identifiable Information

*: California CCPA
Data Security Management for Hybrid Cloud

Consistency

- Most firms are quite familiar with their on-premises encryption and key management systems, so they often prefer to leverage the same tool and skills across multiple clouds.
- Firms often adopt a “best of breed” cloud approach.

Trust

- Some customers simply do not trust their vendors.

Vendor Lock-in and Migration

- A common concern is vendor lock-in, and an inability to migrate to another cloud service provider.
Weakness of Searchable Encryption

<table>
<thead>
<tr>
<th>Unencrypted Data (Plaintext)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: 123-12-1234</td>
</tr>
<tr>
<td>2: 123-12-1235</td>
</tr>
<tr>
<td>3: 123-12-1234</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Encrypted Data (Ciphertext)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: MjM0MjM0MjM0LTEyMz Q1NjlzNDM0DQo=</td>
</tr>
<tr>
<td>2: VGhpcyBpcyBhIHNhbXB lwaGVydGV4dC43=</td>
</tr>
<tr>
<td>3: MjM0MjM0MjM0LTEyMz Q1NjlzNDM0DQo=</td>
</tr>
</tbody>
</table>

Search is utilized in virtually every application and is critical in a collaborative cloud environment. As mentioned, regular encryption hides data so well that search is not feasible. However, it is possible to efficiently search on encrypted data if one is willing to sacrifice some security. In general, any efficiently searchable encryption algorithm shares a common security weakness: equality of keywords is leaked, making certain statistical attacks possible.
The Landscape
Fraud & Identity Thefts

Source: Adapted from FTC Identity Thefts Report, US Federal Trade Commission (FTC)

Source: Adapted from FTC Identity Thefts Report, US Federal Trade Commission (FTC)
Data Privacy Enforcement
Actions Worldwide

Source: Adapted from The PwC Privacy Policy Database, Data Privacy Enforcement Actions Worldwide reported PwC
The Old Corporate IT Environment

Controlled Data

Managed by Corporate

Shared Responsibility

Cloud

On Premises

Database File

Gateway
- WAF
- Web Server
- Application

Browser
- Client

User

Data scientists

Data engineers
- DevOps

Data engineers
- ML architects

Software engineers

Subject matter expert

Line of business
Study: Top Priorities in 2020

- Cybersecurity: 89%
- Remote Enablement: 82%
- Improve Customer Experience: 52%
- Innovation: 46%
- Move to Cloud: 46%
- Update Legacy Infrastructure: 42%
- Hiring: 15%

Source: Hitachi
Study: Top Priorities in 2020

North America: 88%
EMEA: 8%
APAC: 4%

Titles:
- C-suite: 24%
- VPs: 18%
- Directors: 58%

Company Size:
- Enterprise (10k+ employees): 38%
- Small (<5k employees): 54%
- Midsized (5k-10k employees): 8%
An Increasingly Distributed Environment

More than half of companies have transferred from 50% to 100% of their employees to home offices.

In Corporate Control
- User
- Client
- Browser
- Gateway
- WAF
- Web Server
- Application
- Database
- File

Managed by Corporate

Cloud

Uncontrolled Data

Shadow IT

Subject matter expert
- Data engineers
- Software engineers
- ML architects
- Data scientists
- DevOps
Risks & Control in our New Distributed Environment

- High Data Risk
- Low Control

- In Corporate Control
- Cloud
- Shadow IT

Time
Under control? Is the situation getting worse?

1. How do we control privacy of Test Data? Using Prod Data to meet Their Goals? Outsourced testing?
2. Do we have increasingly less control over distributed data when working from home? Attack Surface increasing?
3. Is compliance under control? Is the situation getting worse?
4. How much is End-point security helping? How can we protect against Supply Chain Attacks? Solarwinds?
Factors Impacting Information Security Functions in Three to Five Years

- IoT and cyber physical systems: 13% (Top 3), 43% (1st Choice), Is this surprising?
- Increasing regulations: 14% (Top 3), 40% (1st Choice)
- Evolving threat landscape: 16% (Top 3), 37% (1st Choice)
- Business leader seeking information security as a competitive edge: 8% (Top 3), 30% (1st Choice)
- DevOps-driven IaaS adoption: 8% (Top 3), 29% (1st Choice)

External Factors (First Choice), 30%

n = 403, All Respondents, Excluding Don’t Know or Refused
Q01: What are your organization’s top three drivers that are likely to impact its information security function and controls in the next three to five years?
Source: Gartner 2020 Security & IAM Solution Adoption Trends Survey

© 2020 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and/or its affiliates.
Do I need to address Ransomware and other Attacks on Data?

How can we protect our Data?

1. Monitor Access
2. Protect Data
3. Offline Backups
4. Test restore procedures
Ransomware and other Breaches on The Rise

Electronic Health Records:

- Healthcare’s attack surface has grown considerably over the last two decades.

Verizon DBIR 2021
### Ransomware


**Figure 18:** Percentage of organizations affected by ransomware in the last 12 months, by industry.

Factors Impacting Information Security Functions in Three to Five Years

- **IoT and cyber physical systems**: 13% (Sum of Top 3), 43% (1st Choice)
- **Increasing regulations**: 14% (Sum of Top 3), 40% (1st Choice)
- **Evolving threat landscape**: 16% (Sum of Top 3), 37% (1st Choice)

- **Business leader seeking information security as a competitive edge**: 8% (Sum of Top 3), 30% (1st Choice)
- **DevOps-driven IaaS adoption**: 8% (Sum of Top 3), 29% (1st Choice)

Source: Gartner 2020 Security & IAM Solution Adoption Trends Survey

n = 403, All Respondents, Excluding Don't Know or Refused.
Q01. What are your organization’s top three drivers that are likely to impact its information security function and controls in the next three to five years?

© 2020 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and/or its affiliates.
Describe your organization's deployment plans for each of the following emerging IT security technologies/architectures.

**Software-defined wide area network (SD-WAN)**
- Currently in production: 51.7%
- Implementation in progress: 30.6%
- Implementation to begin soon: 13.3%
- No plans: 4.4%

**Zero trust network access (ZTNA)**
- Currently in production: 30.2%
- Implementation in progress: 44.3%
- Implementation to begin soon: 16.8%
- No plans: 8.7%

**Secure access service edge (SASE)**
- Currently in production: 39.8%
- Implementation in progress: 34.3%
- Implementation to begin soon: 18.3%
- No plans: 7.6%
Cloud
Multi-cloud Policies & Controls

Applications, Data and Users
- Configuration as code (software)
- Cloud provider infrastructure

Privacy

Security

Regulatory Compliance

On-premise

Active Directory, Load Balancers
- WAF
- SIEM
- Firewall
- Encryption
- Tokenization
- Key Management
- AV - Anti Virus
- Network Sec
- And more

Configuration

Policies

Controls
Organizations migrating workloads to the Cloud

<table>
<thead>
<tr>
<th></th>
<th>Biz Apps</th>
<th>Data Warehouses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Currently Migrating</td>
<td>41%</td>
<td>30%</td>
</tr>
<tr>
<td>Migrating in 12 mo</td>
<td>17%</td>
<td>15%</td>
</tr>
<tr>
<td>Migrating in 1-3 years</td>
<td>14%</td>
<td>15%</td>
</tr>
<tr>
<td>% of the market moved or moving with 3 years</td>
<td>72%</td>
<td>60%</td>
</tr>
</tbody>
</table>

The 2020 IDG Cloud Computing Survey
(Represents the 551 IT decision-makers)
<table>
<thead>
<tr>
<th>Area</th>
<th>Timing</th>
<th>Focus</th>
<th>Comments</th>
<th>Use case: Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements</td>
<td>Short</td>
<td>Internal requirements</td>
<td>International regulations</td>
<td></td>
</tr>
<tr>
<td>Cloud</td>
<td>Short</td>
<td>Machine Learning</td>
<td>Start with basic ML training and inference on sensitive data in cloud</td>
<td></td>
</tr>
<tr>
<td>Competition</td>
<td>Short</td>
<td>Competitive advantage</td>
<td>ML and NLP-powered services can give banks a competitive edge</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Short</td>
<td>Encrypted data</td>
<td>Important</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>Synthetic data</td>
<td>Computing cost?</td>
<td></td>
</tr>
<tr>
<td>Analytics</td>
<td>Medium</td>
<td>AML / KYC</td>
<td>What are other Large banks doing?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Short</td>
<td>Analytics</td>
<td>Initial focus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Short</td>
<td>Operation on encrypted data</td>
<td>Computation on sensitive data to the cloud. Trade-offs with performance, protection and utility?</td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>Short</td>
<td>Industry dialog</td>
<td>Working groups in standard bodies (ANSI X9, Cloud Security Alliance, Homomorphic Encryption Org)</td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Short</td>
<td>Encrypted model</td>
<td>Important</td>
<td></td>
</tr>
<tr>
<td>Pilot</td>
<td>Short</td>
<td>Experimentation</td>
<td>What are other Large banks doing?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Short</td>
<td>Scotia Bank case study</td>
<td>Query solution for AML / KYC</td>
<td></td>
</tr>
<tr>
<td>Proven</td>
<td>Medium</td>
<td>Fast follower</td>
<td>What are some proven solutions?</td>
<td></td>
</tr>
<tr>
<td>Quantum</td>
<td>Short</td>
<td>Homomorphic Encryption post-</td>
<td>Lattice-based cryptography is a promising post-quantum cryptography family, both in terms of foundational properties as well as its application to both traditional and homomorphic encryption</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>Quantum</td>
<td>Plan for quantum safe algorithms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>Quantum</td>
<td>Plan for quantum ML algorithms</td>
<td></td>
</tr>
<tr>
<td>Sharing</td>
<td>Short</td>
<td>Secure Multi-party Computing (SMPC)</td>
<td>Without revealing their own private inputs and outputs. Encrypted data and encryption keys never commingled while computation on the encrypted data is occurring or an encryption key is split into shares</td>
<td></td>
</tr>
<tr>
<td>Solutions</td>
<td>Short</td>
<td>Vendor positioning</td>
<td>Nonlinear ML regression needed? Linear Regression is one of the fundamental supervised-ML. Linear and non-linear credit scoring by combining logistic regression and support vector machines</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Short</td>
<td>Framework integration</td>
<td>Important</td>
<td></td>
</tr>
<tr>
<td>3rd party</td>
<td>Long</td>
<td>3rd party integration</td>
<td>Mining first</td>
<td></td>
</tr>
<tr>
<td>Training ML</td>
<td>Long</td>
<td>Federated learning</td>
<td>Complicated</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>TEE</td>
<td>Emerging</td>
<td></td>
</tr>
</tbody>
</table>