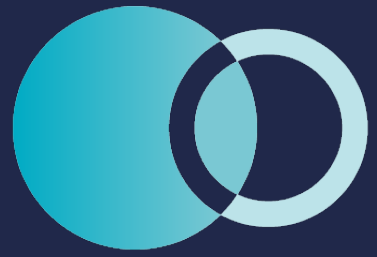
GDPR and Beyond

Demystifying Data Governance Challenges.



agilelab

www.agilelab.it

About us

Francesco Valentini
Data Architect with several years of hands-on experience in the dynamic realm of data management.

Antonio Murgia
Data Architect and Data Engineer specializing in high volume, high throughput, batch and streaming, analytical distributed systems.

Agile Lab

- Italian consulting company specialized in data management
- Fortune 500 customers
- Multiple business units
- Holacracy inspired organization

Agenda

- Data and GDPR
- Anonymization
- Encryption
- Comparison between different strategies
- A viable data-sharing strategy

agilelab

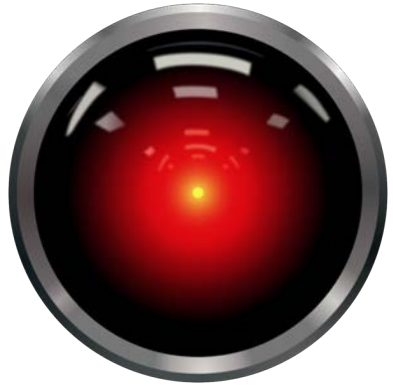
Data is the new oil

Data is the new oil: the good

Data as the Fuel for Innovation

ML

Analytics



AI

Data is the new oil: the bad

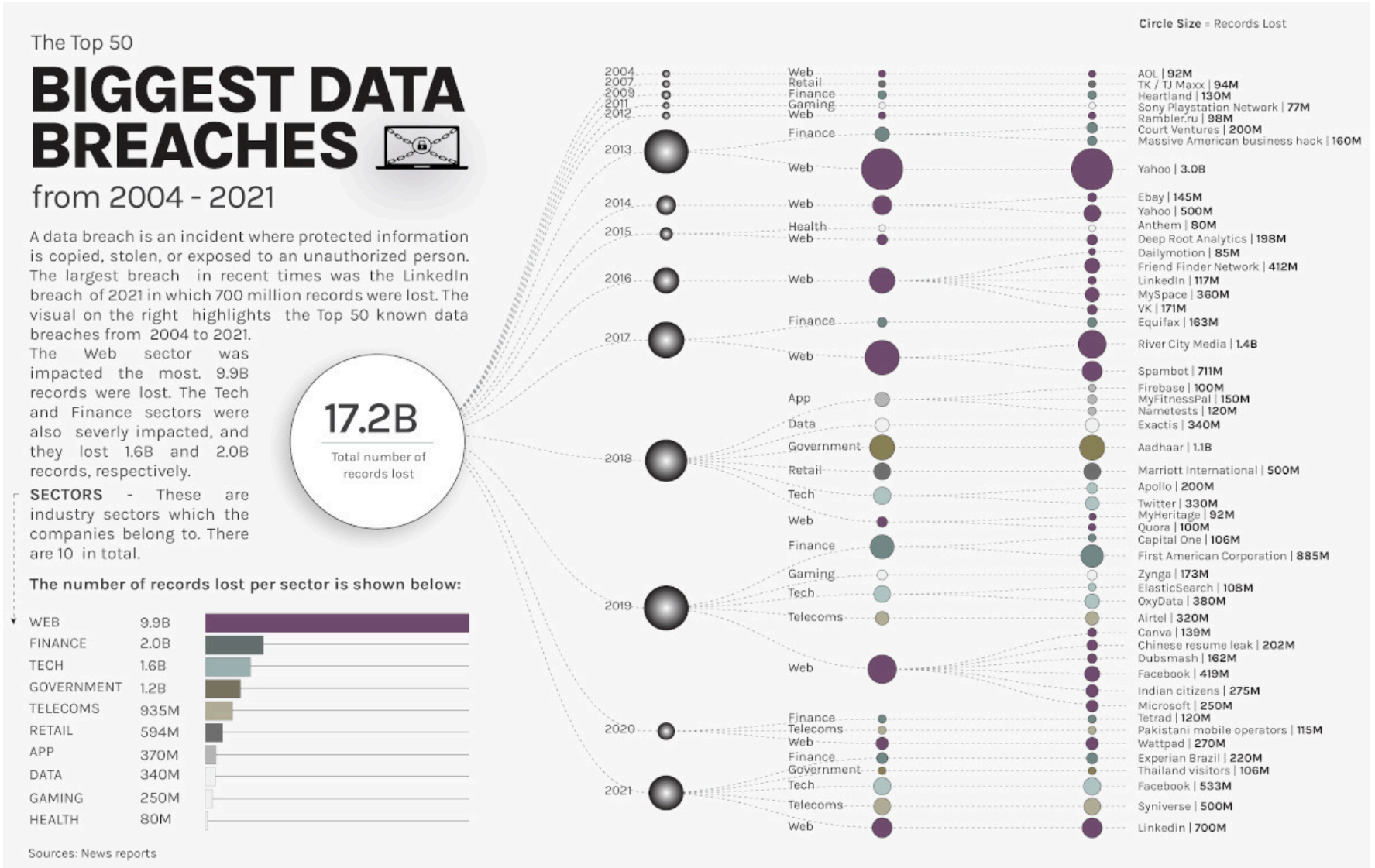
Data Breaches

Regulatory fines

Privacy violations

Reputation Damage

Major data breaches



5B+
GDPR fines

Source: https://www.enforcementtracker.com

agilelab

What is GDPR

25th May 2018

General
Data
Protection
Regulation

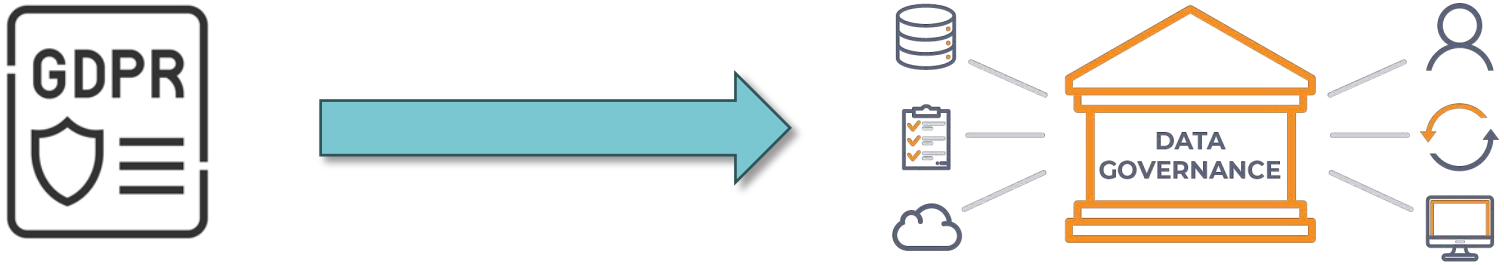
Key principles

- Lawfulness, fairness, and transparency
- Purpose limitation
- Data minimization
- Accuracy
- Storage limitation
- Integrity and confidentiality
- Accountability

GDPR Compliance requirements

- ❑Data protection impact assessments (DPIAs)
- ❑Data breach notifications
- ❑Appointment of Data Protection Officer (DPO)
- ❑Implementation of data protection by design and by default
- ❑Record-keeping obligations

Implications for Data Governance



Data minimization principle

Art. 5 - Par. 1c

Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')

Looks familiar?

Anonymization techniques

Anonymization methods and techniques

- ❑ Anonymization
- ❑ Pseudo-anonymization

Anonymization methods and techniques

Generalization

Name	Age	Birth date	State	Disease
Mark	30	1993-10-19	Texas	Cancer
John	24	1993-06-20	Colorado	Viral infection
Lukas	28	1993-04-11	California	TB
Paul	27	1993-10-19	Florida	No illness

Name	Age	Birth date	State	Disease
Mark	20 < Age ≤ 30	1993	Texas	Cancer
John	20 < Age ≤ 30	1993	Colorado	Viral infection
Lukas	20 < Age ≤ 30	1993	California	TB
Paul	20 < Age ≤ 30	1993	Florida	No illness

Anonymization methods and techniques

Randomization

Name	Age	Birth date	State	Disease
Mark	30	1993-10-19	Texas	Cancer
John	24	1993-06-20	Colorado	Viral infection

Name	Age	Birth date	State	Disease
Mark	35	1993-10-19	Texas	Cancer
John	29	1993-06-20	Colorado	Viral infection

Noise Addition

Name	Age	Birth date	State	Disease
Mark	30	1993-10-19	Texas	Cancer
John	24	1993-06-20	Colorado	Viral infection

Name	Age	Birth date	State	Disease
Mark	24	1993-10-19	Texas	Cancer
John	30	1993-06-20	Colorado	Viral infection

Shuffling

agilelab

Anonymization methods and techniques

Suppression and redaction

Name	Age	Birth date	State	Disease
Mark	30	1993-10-19	Texas	Cancer
John	24	1993-06-20	Colorado	Viral infection

Name	Age	Birth date	State	Disease
Mark	***	1993-10-19	Texas	Cancer
John	***	1993-06-20	Colorado	Viral infection

Suppression

Name	Age	Birth date	State	Disease
Mark	30	1993-10-19	Texas	Cancer
John	24	1993-06-20	Colorado	Viral infection

Name	Age	Birth date	State	Disease
Mark	2*	1993-10-19	Texas	Cancer
John	3*	1993-06-20	Colorado	Viral infection

Redaction

agilelab

Anonymization methods and techniques

Comparison matrix of anonymization tecniques

	Suppression	Redaction	Generalization	Shuffling	Noise Addition
Secrecy	Best	Good	Poor	Poor	Best
Privacy	Best	Fair	Fair	Fair	Best
Utility	Poor	Fair	Good	Fair	Poor

Encryption techniques

Encryption methods and techniques

Format Preserving Encryption

Format Preserving Encryption, or FPE, is a **symmetric encryption algorithm** which preserves the format of the information while it is being encrypted. FPE is weaker than standard *Advanced Encryption Standard (AES),* but FPE can preserve the length of the data as well as its format.

	Credit Card number	SSN	Phone number
Plaintext	4287 1214 5091	101-01-5586	813-204-9012
Ciphertext	3259 1112 4984	304-34-9403	453-112-3838

Tools and implementations

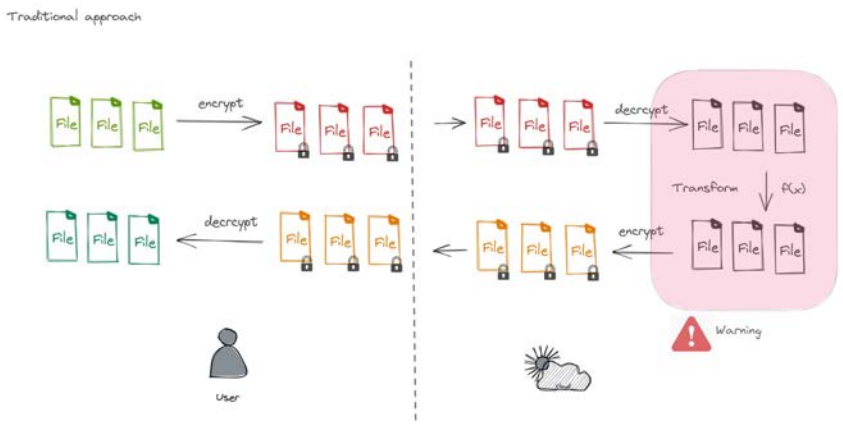
- https://github.com/googleapis/java-dlp
- https://github.com/bcgit/bc-java

Encryption methods and techniques

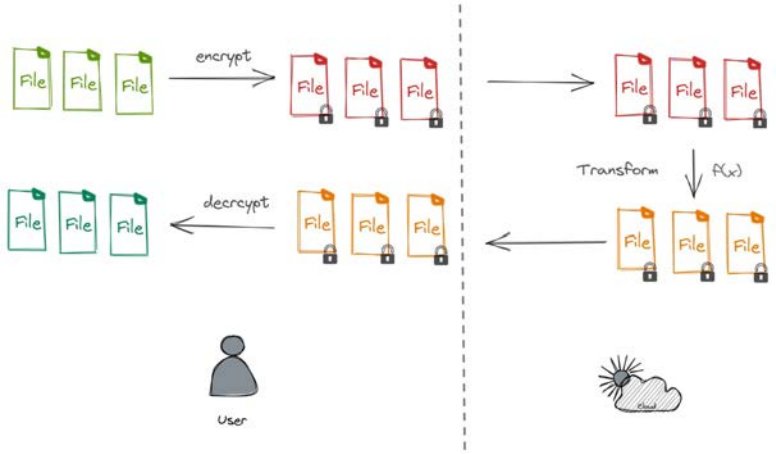
Homomorphic Encryption

- **Partially Homomorphic Encryption**
- **Somewhat Homomorphic Encryption**
- **Fully Homomorphic Encryption**

Before



After



Notable implementations:
Microsoft SEAL, PALISADE, HEAT, NFLlib, Concrete , FHEW / TFHE

agilelab

Other techniques...

Other techniques

- Tokenization

- Synthetic Data Generator

Brief recap

Sampled data

- Realism/Relevance
- Improved testing
- Stakeholder confidence

- Privacy and compliance
- Freshness
- Security

Synthetic data

- Privacy and security
- Availability
- Efficiency

- Lack of realism
- Overfitting
- Validation challenges
- High setup complexity

Encrypted data

- Privacy Protection
- Regulatory Compliance
- Data Sharing
- Risk Mitigation

- Complexity in Data Use
- Key Management
- Limited Testing Accuracy

Anonymized data

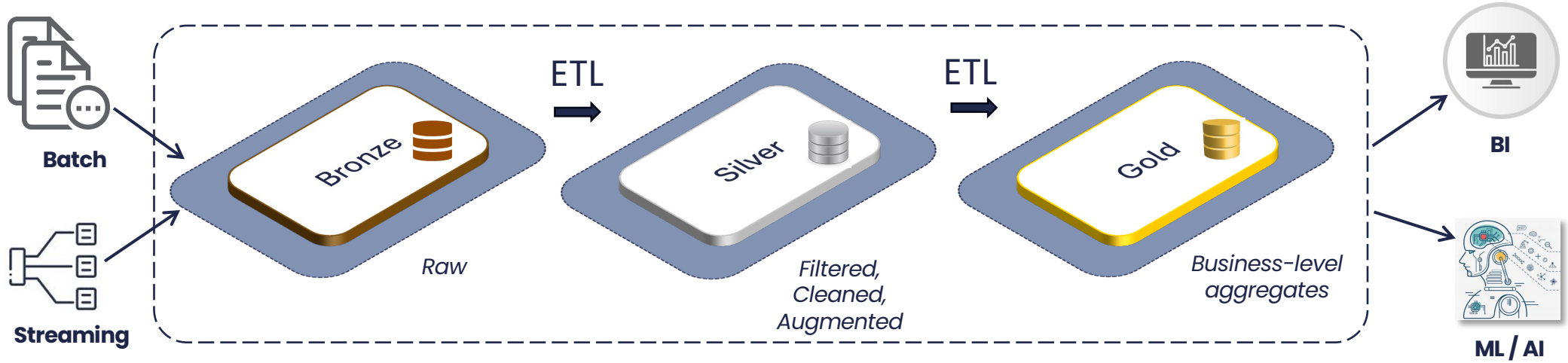
- Privacy Protection
- Regulatory Compliance
- Data Sharing
- Risk Mitigation

- Complex Anonymization Process
- Data Duplication
- Re-Identification Risk

A viable data sharing strategy

Secure data sharing practice

Quick overview on Medallion Architecture



agilelab

Secure data sharing practice

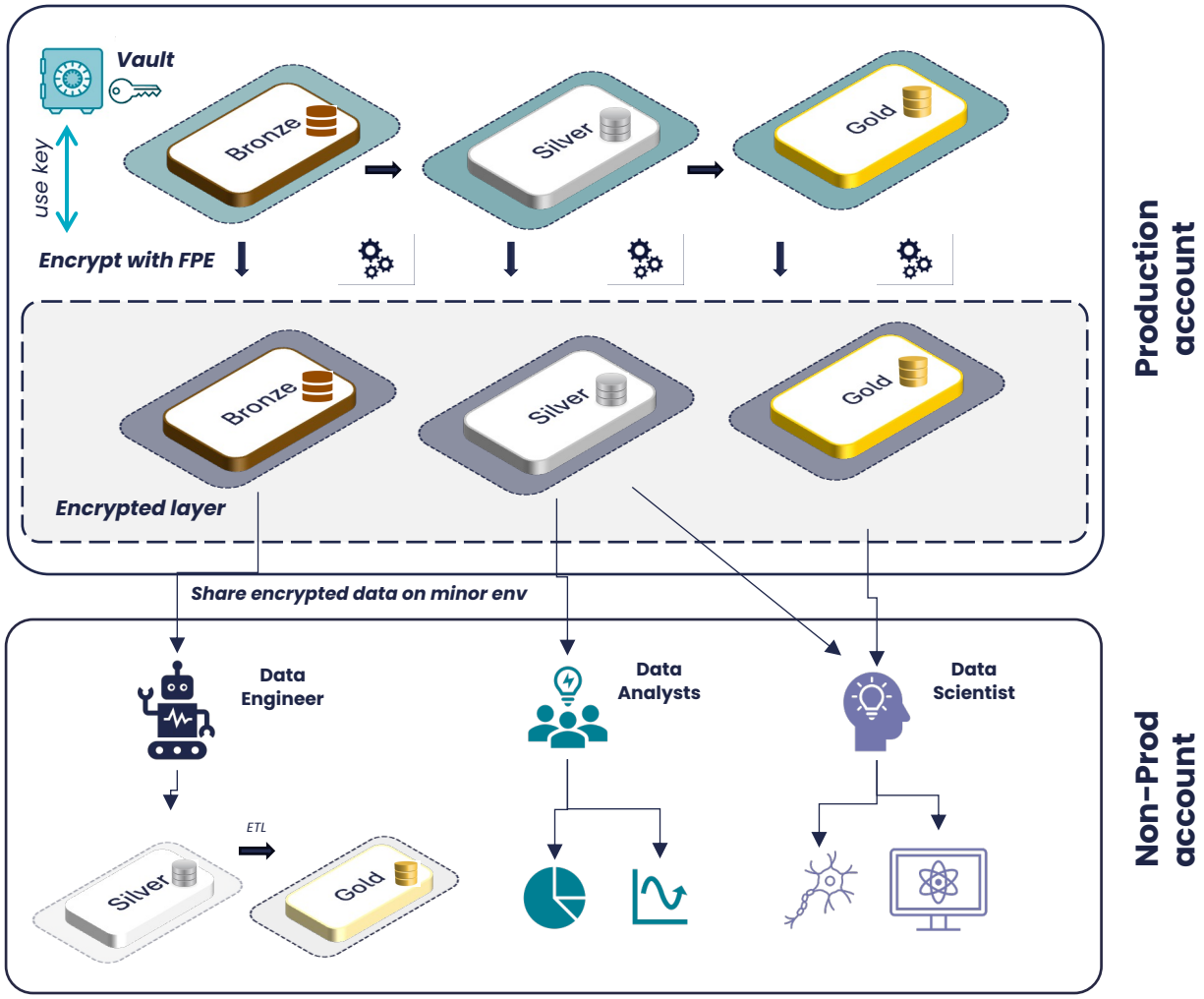
Recipe for a cloud based scenario eg. AWS

Step 1

- Data teams are in charge to productize the job and anonymize data.
- The encryption process becomes a mandatory step in the data life cycle made of data ingestion, data normalization, data harmonization and delivery.

Step 2

- Open a read-only cross policy account from prod to non-prod.
- The minor environment never writes to prod, it is only enabled for reading operation.
- Encryption key is never shared with minor environments



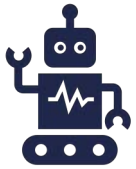
agilelab

Secure data sharing practice

Benefits

Format preserving encryption guarantees references integrity, no schema changes across different datasets and allows to re-use business logic without code changes

Dev applications are allowed to read only the encrypted layers leveraging ad-hoc IAM policy. This simplify the data movement and the orchestration process between environments



ML Engineers and data scientist can prototype and train their models on “quite real” data on a safe layer reducing the risk of inconsistent performance when moving to prd

DevOps practice is still in place since deployments of new artifacts and model can follow the standard CI/CD flow across multiple env: DEV -> QA -> PRD

Minimization principle is respected on minor environments since we do not have sensitive and PII data so the attack surface is strongly reduced.

agilelab

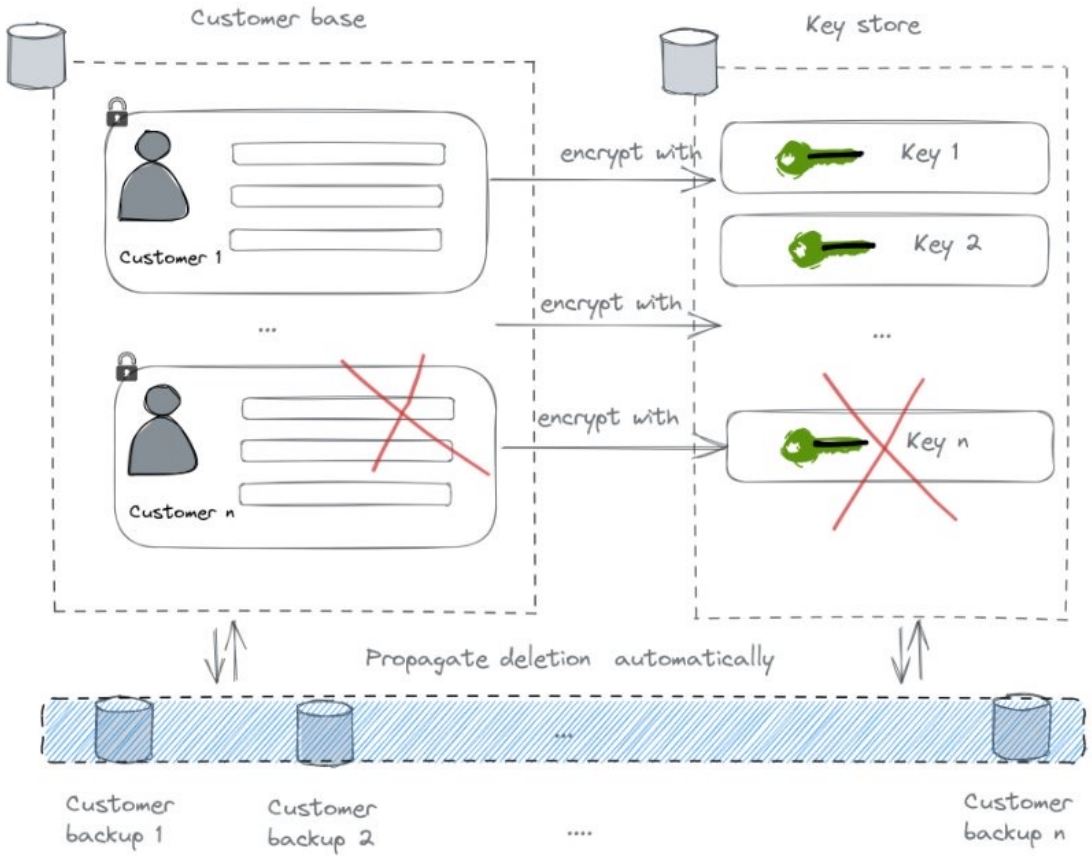
Add-on

Right to be forgotten

Crypto shredding

Crypto-shredding is the practice of 'deleting' data by deliberately deleting or overwriting the encryption keys. This requires that the data have been encrypted.

Deleting the key will automatically logically delete the record on all the existing copy, since all the encrypted info are not reversible anymore.



agilelab

agilelab

Thank you!

www.agilelab.it

@agilelab_official

linkedin.com/company/agile-lab

@agilelabsrl