Thinking with a *data protection* mindset

Achieving data security readiness in the age of generative AI and post-quantum computing requires a fundamentally disciplined approach

Capgemini

In the digital world, data is the basis of programming languages and operating systems, sound and images, emoticons, medical records, and identities. In 2020, the world produced some 59 zettabytes (59m petabytes) of data, predicted to become 175 zettabytes by 2025. Data has become valuable – and often essential – to individuals and the organizations that serve them. This necessitates careful stewardship, secure storage, and protection. Governments and regulators around the world have acknowledged the importance of this issue by enacting legislation to ensure its governance, protection, and proper usage.

The European Union's General Data Protection Regulation (GDPR), enacted in 2016, along with the dawning of data's commercial value around the same time, opened the chapter on structured corporate data management. GDPR mandates personal data to be processed lawfully and transparently, used only for specific purposes, retained only as long as necessary, and protected with appropriate safeguards.

Despite efforts to instill rigorous processes, data breaches are inevitable. The complexity of systems, technology, and humanity means that the risk of breaches can be reduced and managed but not eliminated. Even standards of encryption currently thought to be invincible may have an as-yet-unknown vulnerability, as in the case of RSA private keys, a commonly used type of encryption, found to be vulnerable in some cases due to ***hardware faults.*** Human ingenuity is often matched by acts of malice or simple mistakes.

The nine essential components of a data protection framework

Managing an organization's data protection for the long term requires a specific structure to plan for its expansion, location, lifecycle, and the allocation of responsibility for its long-term guardianship. An organization undertaking data transformation should use a cross-disciplinary approach to include governance specialists, business process experts, data stewards, and data engineers.

The following nine-point framework coordinates people, policies, processes, strategies, standards, and technologies to allow an organization to use data as a critical business asset. It should make an organization's data consistent, reliable, secure, and auditable throughout its entire life cycle.

1. **Governance** establishes data as a business asset. It ensures a sustainable means of utilizing data to achieve the organization's business goals and purposes. IT and business collaborate to define the rules and strategies governing data and its elements, from acquisition, management and storage, to utilization and visualization. It defines its ownership and communication protocols between IT and business experts, sets the data organization and operating model, enforces data policies and processes, and promotes company data literacy and culture.
2. **Discovery** of an organization's data assets (in data repositories, domains, lakes, and warehouses) determines the data entities and their attributes, as well as their lineage and traceability. It involves identifying sensitive data using processes of data discovery, classification, and risk analysis. It is at this point that data profiles are defined with support from business specialists.
3. **Data glossary:** this is a compilation of semantic descriptions of enterprise-wide business concepts and data standards. They are usually business term definitions for critical data attributes and their relationships to each other, with examples, exceptions, synonyms, lineage, business rules, context notes, and data ownership.
4. **Cataloging:** recording data attributes, descriptions, and locations for an inventory of data assets through the discovery, description, and organization of distributed datasets. This can be within systems, databases, and data lakes. It enables data users to find and understand relevant datasets.
5. **Data quality** is assessed during data profiling by measuring and visualizing it across a range of quality dimensions; data hygiene (cleansing, enriching, or other improvement of data) and functionality are defined here, with a focus on setting strategic and operational priority metrics.

6. **Master Data Management (MDM)** sets a "single source of truth" for data. This is a single view of master and reference data based on a variety of source systems that all hold similar, though incomplete, master data. It sets an authoritative source for data restoration from backups in case of loss or damage.
7. **Data Lifecycle Management** the central point for data creation and destruction. An enterprise manages its data assets lifecycle via data retention, archiving, and disposal policies. Relevant timings, methods, target repositories, and compliance requirements are recorded here, with specialized data management tools configured to perform recurring or one-off lifecycle management tasks. Solutions should be capable of data discovery, classification, and security, with the ability to mask or delete data, typically via MDM.
8. **Data Mesh Governance** is an architectural and organizational paradigm that aims to decentralize and distribute ownership and responsibility for data delivery to the people closest to it. It provides a standardized set of data products for self-serve data consumption, often via data marketplaces. The outcome is data repositories with a separation of data types. Storing data without proper governance has resulted in costly, unmanageable multipurpose data lakes with a jumble of expiry dates. If not controlled, they pose significant compliance risks.
9. **Data ontology** follows from the previous step and applies across the business. It can be an enormous undertaking since it requires interpretation of relations between data from many departments.

Putting the framework into action

Those responsible for influencing an enterprise's data should consider two questions:

- What do regulations allow the company to do with the data?
- What activity is compatible with the business's strategic objectives?

Fundamentally, the answers to these questions are to be found in a company's data governance policies. For example, a customer may want to apply machine learning to data that is sensitive intellectual property. This would require it to be pushed to the cloud for processing, but such usage could be forbidden by regulations, notably in defense, financial, and insurance sectors, by a company's own data security policy, or by data sovereignty rules. Overcoming these hurdles may be possible with an appropriate data encryption methodology in harmony with regulations and a clearly defined data lifecycle.

Data encryption is an essential aspect of implementing the framework. It is a combination of an encryption algorithm and an encryption key. The key can be managed on-premises, in the cloud, or by an outsourced service. Since the algorithm is owned by the technology provider, a user has the most control over key management, which is kept separate from certificate management. The whole encryption key lifecycle can be managed from a single location. The key is generated by a hardware security module, which should be owned by the customer.

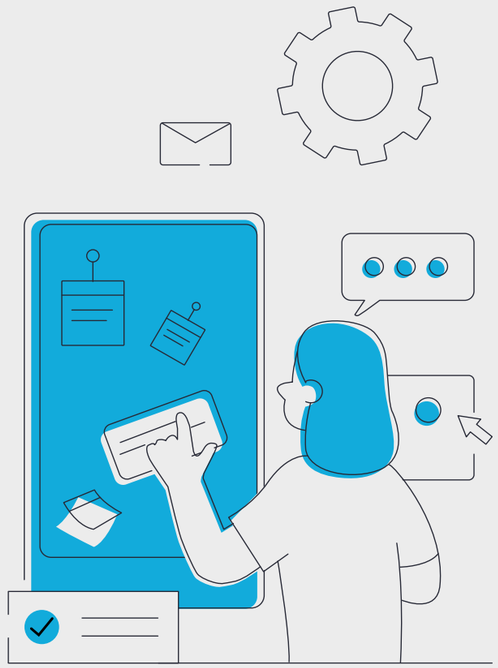
Applying a risk-based approach, the process can be stress-tested in a proof-of-concept (POC) phase with a technology partner, such as a hyperscaler or a post-quantum computing lab, to meet security requirements. The POC can be run securely with candidate algorithms, which may not yet have been validated.

Hardware security modules for automated key and certificate management

A Hardware Security Module (HSM) is an essential part of an automated key and certificate management system. It secures critical data and digital identities by generating, managing, and storing cryptographic keys and certificates. Relying on a third-party HSM provider allows you to generate new keys, which are stored and backed-up automatically.

Using an outsourced data security provider to manage the HSM is also a potential answer to the industry's persistent short supply of cybersecurity specialists. There is also the potential benefit that, by lowering business risk with a structured, externally-validated automated approach, the cost of insurance is reduced.

Automation and streamlining in key and certificate management, as well as acting as safety nets for renewals and backups, also cut time-to-market in software development, security, and operations (DevSecOps) processes.



Life sciences – lifting and shifting securely to the cloud

A large life sciences company headquartered in Europe wanted to reduce overhead by closing some of its less efficient data centers and moving to the cloud. The first phase of migration involved re-hosting more than 450 virtual machines (VMs) in the public cloud. In doing so, the company had to comply with the data privacy and data sovereignty regulations in GDPR and the European Court of Justice's Schrems II judgment, which had implications for the transfer of personal data beyond the EU.

The VMs were encrypted natively on-premises with encryption keys kept separately in a hardware security module (HSM). They were then moved to AWS, the cloud service provider. AWS had no access to the VMs' encrypted data while this was being done.

The HSM was deployed to manage encryption keys in the cloud and protect sensitive assets from access by server provider employees and unauthorized users. This ensured that neither AWS employees nor external attackers could have access to encryption keys or sensitive data.

The project established rigorous data sovereignty and privacy compliance in the cloud with a strict separation of duties between AWS and the company's security team. It also brought the benefits of automated key and certificate management.

Automation belongs with encryption policy

By not managing all encryption keys in a unified, automated manner, a significant systemic malfunction is likely for many organizations.

The following scenario is not rare: a key is generated with a multi-year lifecycle, according to best practice. However, when it is due for renewal, the specific knowledge of the key's properties, application, and environment could remain with those responsible for its initial maintenance, who may have changed roles or left the company. This may not emerge until the key expires and service is disrupted. More disastrously, all keys are stored on a misconfigured cloud server, which becomes publicly accessible so that all encryption keys are leaked. Affected applications could be in use for ten years or more, causing a mismatch in lifecycles.

Automation is the essential solution to efficient, consistent key management. Cloud computing requires up-to-date security standards across multiple regions. Each newly generated environment, i.e., storage for a new version of an application deployed by the continuous integration (CI) and continuous delivery (CD) pipeline, requires a new key. Attempting to manage certificate renewal manually – e.g., in a scenario with 100 applications – is highly risky. Automation securely and reliably executes key creation, renewal, backup (including for future legal and regulatory requests), and destruction.

What generative AI means for data security

The commercial possibilities of generative AI have already led to greater demand for data, adding to the argument for a comprehensive data security framework. The greatest risk to data raised by Gen AI is data exfiltration, either intentional or accidental. Prompts can be used to retrain a large language model's (LLM) inference model or leak data accidentally. Data governance sets policies as guardrails to delimit how an LLM can use an organization's data sets.

A cautious approach to data security with generative AI involves defining and locating sensitive data. This highlights the importance of its prior classification and identification so that it can be isolated and blocked from propagation by shadow AI. Creating a golden circle of data security is technically easier with the adoption of deperimeterization, i.e., with zero trust architecture, which moves protection closer to assets, including data.

Capgemini's data protection experience

We see data security as fundamental to safely and sustainably delivering the innovations that could help transform society for the better. We are located in more than 50 countries with a network of connected cyber defense centers and have experience in delivering cybersecurity transformation for private and public sector organizations and in all industries.

Experts to contact

Jérôme Desbonnet
VP - Cybersecurity Chief Technology Innovation Officer
jerome.desbonnet@capgemini.com

About Capgemini

Capgemini is a global business and technology transformation partner, helping organizations to accelerate their dual transition to a digital and sustainable world, while creating tangible impact for enterprises and society. It is a responsible and diverse group of 340,000 team members in more than 50 countries. With its strong over 55-year heritage, Capgemini is trusted by its clients to unlock the value of technology to address the entire breadth of their business needs. It delivers end-to-end services and solutions leveraging strengths from strategy and design to engineering, all fueled by its market leading capabilities in AI, cloud and data, combined with its deep industry expertise and partner ecosystem. The Group reported 2023 global revenues of €22.5 billion.

Get the future you want | www.capgemini.com

GIS_042024_SC