Privacy Compliance for Business Data

Developments on legal and technical fronts to prioritize consumers' personal data protection have produced a lot of discussions, regulations, and documents for privacy consent of end users. But the internal handling and processing of personal data have largely remained untouched.

There is still a scarcity of documentation available on securing the privacy of business-sensitive data that is used internally within businesses. This data is used by data engineers to build data sets, by data scientists for ML modeling purposes, and by data analysts for analytics reporting purposes.

GDPR entitles data subjects to exercise their rights. Businesses can enable the exercise of data subject rights by locating and accessing all data that relates to an individual. However, user-specific, business-sensitive data does not reside in front of or within silos. Instead, it exists across multiple locations, often unclassified, within aggregated datasets, and with varying sensitivities.

That's where the concept of data ownership comes into action, requiring a range of technical, organizational, and procedural controls for organizations to introduce for effective compliance practices. Compliance management for business data also includes implementing strong security measures in place, developing governing policies and procedures, managing access, and other key requirements specific to data protection law of a jurisdiction.

Complexities with compliance

I. Governance policies for different categories of sensitive data may differ.

The scope of personal data transcends typical identifiers like name, identification number, biometric data, etc. to cover an individual's preferences and demographic data, among others. Based on their sensitivity and the risks associated with storage and use, different categories of personal data require varied levels of protection and retention policies.

For example, medical and financial data comprise the most sensitive data. Other data types, such as personal identification data, require access to authorized persons only. In addition to robust security measures, some data types also require special data governance policies and procedures. For example, health data, financial data, biometric data, and geolocation data requires strict limitations and compliance with respective regulations.

II. Application of data science on datasets can reveal sensitive data.

Data science involves analyzing and processing a large amount of data. Techniques used in data science, such as linkage or combining multiple datasets, can potentially uncover patterns, revealing sensitive information.

- Data analytics or machine learning techniques can reveal correlations between seemingly unrelated data points, which may be indicative of sensitive information. For instance, while analyzing a dataset that contains information like individuals' web browsing

history or medical data, a data scientist may discover that individuals who browse for specific medical conditions are likely to have a particular health condition. Such correlations potentially reveal sensitive health information.

- Inferences are widely applicable in personalized advertising, fraud detection, and risk assessment. Inferences can be made by training machine learning algorithms to recognize patterns in the available data and using those patterns to make predictions about certain data attributes of individuals. For instance, banks or credit card companies could train an algorithm to detect and prevent fraudulent spending patterns like purchases made in multiple countries within a short time.
- Even after anonymization, machine learning models can be manipulated to re-identify or de-anonymize data. The presence of enough unique identifiers or a combination of datasets with other sources of information enables re-identification. For instance, in 2008, two researchers using statistical techniques matched a supposedly anonymous Netflix Prize dataset containing movie ratings to specific IMDB movie ratings that were publicly available, re-identifying individual users and inferring sensitive information about their movie preferences.

III. Data sharing across organizations for computational or research purposes risk data loss or exposure.

Data sharing across different organizations becomes necessary to access larger or more diverse datasets for collaborative research, business partnerships, industry analysis, data mining, and emergency response purposes. Such scenarios give rise to the risk of unauthorized access, data breaches, or loss of data control or ownership, especially when multiple parties do not trust each other or when the sensitivity of the data doesn't allow for free access or sharing. To ensure safe data sharing practices, appropriate security measures in addition to data sharing agreements have become the need of the hour.

IV. Data aggregated from different sources make it challenging to attribute ownership of specific data points to individual users.

Data gathered from distinct sources for business analytics can be hard to trace their origin and, therefore, challenging to attribute ownership of specific data points to individual users. For example, a healthcare organization may aggregate electronic health records, bills, and patient feedback to gain insights into patient satisfaction, but enabling the exercise of data subject rights may not be practical. Such instances require keeping tags on anonymized data or making use of AI-powered Data Subject Access Request (DSAR) solutions to provide users with greater transparency and control over their data.

Simplifying compliance for business data

I. Data sources and types identification

The way to compliance is anything but straightforward. It starts with identifying data sources where data may be located in structured or unstructured formats. Unstructured data is typically found in emails, social media feeds, documents, and multimedia files. Techniques like machine learning algorithm training, natural language processing (NLP), named entity recognition

(NER), data discovery tools, and keyword matching help extract sensitive data from unstructured datasets.

Structured data includes financial data, sales data, user data, etc. These are normally stored in databases, data lakes, data warehouses, application logs, and cloud services. Extracting sensitive data from structured data encompasses identifying the scope of the sensitive data elements as per the applicable governing laws or regulations, developing extraction rules like regular expressions and string matching, and securely storing the data in encrypted or masked formats.

II. Charting a data flow map

Data mapping is a recognized way of tracing the digital railroad of data used within a business' IT infrastructure. It catalogs sensitive information from its origin to its transit through and beyond the organization to where it is stored. The overlay captures all stages in the life cycle of sensitive information, including protocols, encryption status, retention policies, access controls, etc. Data flow maps also help businesses become GDPR compliant. It proves beneficial in keeping records of processing activities (Article 30), performing DPIAs (Article 35), demonstrating privacy by design (Article 5), establishing a lawful basis for processing (Article 6), detailing data practices (Article 12), and managing data subject access requests (Articles 15–18, 20–21).

1. Determine the data flow
   - Identify the types of personal data the business collects, including customer data, employee data, vendor data, partner data, or any individual that interacts with the business.
   - Identify the methods of collecting data. Its origin may be through cookies, social media accounts, online forms, paper forms, in-person interactions, phone calls, etc.
   - Identify data processing operations and activities. The data may be sorted, analyzed, filtered, transformed, or mixed with other data through the use of automated tools or manually.
   - Identify where personal data is stored. Data storage systems comprise CRM systems, company servers, local machines, paper records, databases, cloud-based storage systems, etc., including backups and archives.

2. Determine the data access
   - Identify individuals with access to personal data, such as IT administrators, data processors, customer care representatives, marketing personnel, contractors, and third-party service providers.
   - Identify their roles and responsibilities. It assists in determining the job functions and tasks performed by specific individuals, ensuring that access privileges are granted only to those with legitimate needs.
   - Implement access-based controls.
     - ❖ Develop governing guidelines and rules: Policies establish workflows for approving access requests, empowering businesses to identify who has access privileges, monitor the access, and revoke access when no longer required.

- ❖ Access control models: Role-based access controls (RBAC) and attribute-based access controls (ABAC) ensure only authorized personnel get access to personal data. RBAC ties access privileges to jobs and responsibilities, whereas ABAC focuses on attributes associated with the user (department, security clearance level, role, group), the resources being accessed (sensitivity, type, ownership), and the environment in which access is requested (device type, geographic location of the user, network security level).
- ❖ Authentication and authorization mechanisms: Authorization mechanisms involve setting up access permissions and configuring systems to restrict permissions based on users' roles and attributes of the resources. Authentication mechanisms involve passwords, biometric authentication, or multi-factor authentication to verify the identity of individuals attempting to access personal data.

3. Map the flow

Data flow diagrams or flowcharts provide a visual representation of the data flow map, pinpointing information being processed, systems involved in processing, and individuals responsible for processing. In addition to ensuring compliance with data protection laws, it helps organizations identify potential risks and vulnerabilities in their data handling processes for improving data protection measures.

III. Implementing data governance policies and procedures

Access to sensitive data by stakeholders is important for many reasons. For example, to train an ML model for the purposes of weighing customer feedback on a product, data engineers need access to user-specific information, such as delivery addresses and purchase histories. Simply handing over critical information to data engineers or analysts can be risky or substandard. Ideally, first the characteristics or potential risks of the dataset should be explored to identify issues that could lead to data exposure or breach.

A range of technical, legal, or ethical considerations are required to protect the personal data of individuals represented in the data. Implementing data governance policies and procedures embodies the following steps:

a. Data classification

Data classification helps organizations identify and protect their most sensitive data. Data can be classified based on various factors, namely level of availability (e.g., emergency response data, public-facing data), level of sensitivity of the data (e.g., confidential, private, public), level of integrity (e.g., financial data, legal documents), the legal requirements governing the data (e.g., HIPAA, GDPR, PCI-DSS), and the impact on the organization's operations and reputation in case of a breach (e.g., customer data).

Once data is appropriately classified, it can be labeled with designations like top secret, secret, confidential, or restricted to ensure appropriate handling of the data and no risk of unauthorized disclosure. Post-tagging, appropriate security measures—physical or technical—can be applied.

b. Data encryption

Data engineers and scientists should also ensure that the data sets are properly anonymized or de-identified before running any processing mechanisms. Data encryption converts sensitive information into an unreadable format. However, its inability to modify the underlying data itself allows authorized users (data scientists, analysts, and engineers) with appropriate keys to identify individuals from the encrypted dataset. Subsequently, the use of additional techniques anonymizes personal data while still allowing for meaningful analysis of the data.

These techniques are:

- ★ Masking: This technique hides specific data elements of personal information by replacing some or all characters with a symbol, such as an asterisk (*), without altering the structure or format of the data. It is useful for hiding certain characters or digits in a data field, such as social security numbers, credit card numbers, phone numbers, names, or addresses.
- ★ Generalization: This technique is used to obscure specific details in a dataset that are too specific or detailed to identify individuals in the dataset. It is achieved by diminishing the level of detail or precision. For example, a range of ages (e.g., 40–50 years old) or a larger geographic location (e.g., a state or country) could be used instead of reporting an individual's exact age or location, respectively.
- ★ Suppression: When the presence of certain fields in a dataset makes an individual likely identifiable, the suppression technique ensures anonymity by simply not including certain variables or by removing them from the dataset altogether. To illustrate, if a dataset includes demographic information like age, gender, or address, suppression allows removing any dataset to avoid the risk of identification.

For user's discretion, it is advised not to rely solely on one anonymization technique but to err on the side of caution and use a combination of techniques to ensure the highest possible level of anonymity.

c. Data minimization

Data minimization is a fundamental principle under Article 5(1)(c) of the EU's GDPR. As one of the key data governance policies, it limits an organization's data collection, processing, and retention activities to the extent that they are necessary for a specific purpose and not beyond. Data minimization practices include collecting only the minimum amount of data needed to provide a service, deleting data after use, and anonymizing data where possible.

d. Data retention, disposal, and archiving

Stakeholders' awareness of the data usage policies encourages lawful, fair, and transparent data management. It further ensures that the data is not kept longer than necessary and is properly disposed of after use. Data retention requires businesses to keep personal data for a specified period of time. Data disposal involves permanently deleting electronic or physical records of the data in such a way that they can't be retrieved or reconstructed. Data archiving involves securely retaining such data that is not required for quotidian business operations but may come in handy for auditing, business continuity planning, or regulatory requirements.

Data retention, disposal, and archiving goals could be met with these techniques:

★ Secure data destruction: It involves the use of specialized software to overwrite the data multiple times, rendering it unreadable and unrecoverable. Physical destruction of USB drives or hard drives also makes data irretrievable.

★ Regular data purges: This involves reviewing stored data regularly to delete those that are no longer needed. It minimizes the amount of personal data stored by an organization and reduces the risk of data breaches or unauthorized access.

e. Model validation

It entails regularly validating machine learning models to ensure that they do not reveal sensitive information or perpetuate bias. Validation, through careful analysis of the model's outputs and testing against different datasets or scenarios, can identify if the model is using sensitive features in its predictions or making predictions with a high degree of certainty that could expose personal information. Based on identified vulnerabilities, the model's design could be re-evaluated, its parameters adjusted, the training data updated, and then the model re-validated to ensure ML models are accurate, reliable, free from bias, and protect the privacy and security of individuals and organizations.

f. Privacy-preserving techniques

★ Differential privacy: It first involves computing sensitive analysis of the data by adding or removing a single value and then adding random noise based on the desired level of privacy to mask individual records. This technique provides strong privacy guarantees for individuals while still allowing for accurate analysis and machine learning.

★ Homomorphic encryption: HE requires the generation of a public and a private key to encrypt and decrypt data, respectively. The encrypted data undergoes computation through servers that perform computations using mathematical operations specifically designed for HE data. These operations allow computations to be performed directly on encrypted data without first decrypting it. Once computations are performed, the data owner decrypts the results using the private key.

★ Secure multi-party computation: The use of SMPC is seen in carrying out a joint computation on private data by multiple parties with trust issues or where the data needs to be kept confidential. Similar to homomorphic encryption, each party generates its own public and private keys for encryption and decryption purposes. Once keys are generated, data is partitioned into smaller subsets and distributed among each party, which then compute their own part of the function using their private data. Thus, each party shares encrypted results with other parties to compute their own part of the function without ever seeing the original data. Finally, using their private keys, each party gets to see their final decrypted result.

g. Privacy impact assessment

Article 35 of the GDPR requires organizations to conduct a PIA, especially when the processing of sensitive information is involved and its results can risk the rights and freedoms of data subjects. PIA provides a framework for identifying, assessing, and mitigating privacy risks associated with an organization's products, operations, or services. Conducting a PIA helps

organizations find privacy risks early in the planning process, prompting proactive privacy and security considerations in their operations.

Conclusion

Staying transparent about internal and external data practices helps develop trust with users and ensure compliance with data protection regulations. Getting internal data practices compliant can be tough, but with the above technical, organizational, and procedural controls, the road to compliance becomes short and smooth. Last but not least, regular training and education programs should be provided to all the stakeholders, including data engineers, data scientists, and data analysts, to ensure that they understand their responsibilities for protecting sensitive data and are aware of the latest privacy regulations and best practices.