THE COMPUTERTECH
(An International Peer Review Journal)
VOLUME 8; ISSUE 1 (JAN-JUNE); (2022)
WEBSITE: THE COMPUTERTECH

Automated Data Retention Policy Enforcement in SQL Server Environments: An AI-Governed Lifecycle Management Framework

Divya Sai Jaladi[1*]

[1]Application Developer, SCDMV, Charlotte, NC, UNITED STATES

Abstract

There are a vast number of interconnected transaction level facts that are contained within relational databases. These facts can collectively (or separately) represent an infinite number of business records. We provide a framework in this article that allows for the definition of business records that are stored in a database system as queries, followed by the specification of retention policies which are applied to those searches. When it comes to creating and executing records retention rules for relationally stored documents, we also emphasise many critical challenges that policy makers and database administrators need to take into consideration.

Keywords: Data Retention; Policy Enforcement; SQL Server; Automated Compliance; Data Lifecycle Management; Artificial Intelligence.

Introduction

The legal ramifications of failing to retain corporate data for mandated durations, along with the liabilities linked to excessive preservation, have generated significant interest in automated records retention systems. Regrettably, commercial software solutions for records management have mostly concentrated on the preservation and prompt destruction of company emails and documents.

Conversely, relational databases serve as far more comprehensive data repository than papers. A multitude of essential business records may be generated on demand by accessing the database at any specific moment. Databases are often structured such that a single physically stored piece of information may concurrently belong to several entries across different functional domains of an organization. The volume of information (the quantity of records) that may be extracted from a database is limitless. Numerous records may be indirectly generated by creatively modifying existing data. A database that retains client invoices and purchases may also facilitate the generation of a record of things that a certain customer has not acquired, despite this information not being explicitly recorded inside the system.

Relational Records

Our suggested methodology for records management in relational systems adopts a query-oriented approach to record definition. We contend that if a database is subject to legal discovery and forensic examination, then any information extracted from the data it contains qualifies as a record held by an entity. Consequently, we define a record in a relational database system (relational record) as the output produced by a certain data retrieval query at a particular moment in time. As

not all entries in a database may pertain to policy relevance, each business record subject to a certain retention policy must be expressly delineated as a query. This approach's elegance and flexibility stem from modern query languages, like SQL, which enable users to retrieve individual rows of data and generate intricate records that aggregate information from multiple database segments, thus simplifying the specification of complex records. However, a need for this strategy is that the user (policy maker) must possess a certain level of familiarity with the specific query language used by the database system.

The reader should be aware that our definition of a record in a database system does not delineate the specific contents of a record but rather provides a meta-description of it. A record is defined as a means for obtaining, gathering, and organising particular information recorded in a database at any given moment.

Due to the ongoing change of database contents, the concept of temporality is closely linked to relational records. A temporal query that obtains all information of client invoices generated over the last five days from a database serves as an exemplary illustration. While the concept of the record and its temporal dimension (5 days) are established, the content inside this record is always evolving. Consequently, rules established over such records are not based on permanent data elements but rather on fixed metadata elements that indirectly reference the data that may be deemed relevant to the policy at any given time. The divergence between the concept of a record and the data it encompasses presents a considerable barrier in translating the knowledge of policymakers and compliance officers to the realm of data storage.

In the European Union, Intelligent Infrastructure Management encompasses Smart Infrastructures that integrate many operators across several sectors, including energy, public transit, and public safety. An intelligent infrastructure refers to "(…) an interconnected sensing network that delivers real-time digital information regarding the system's condition 11" 12. Consequently, intelligent buildings possess controls that function autonomously, augmented by sensor-enhanced capabilities to Governance Framework for Intelligent Infrastructure Ecosystems. Legal and regulatory structure. According to Weber and Žarko, regulatory characteristics pertain to legal statutes and regulations pertinent to service providing. They are established by national legislation or EU directives and regulations applicable to all EU member states. The paramount characteristics of a service pertain to: i) lawful interception of data traffic, encompassing IoT traffic; ii) service reliability, specifically the capacity to avert frequent and significant service disruptions; iii) protection of personal data; iv) secure systems that thwart cyber-attacks at both the device and service levels; v) operator transition.

Specifically, the capacity to alter an IoT operator within the value chain; regulated services pertaining to roaming devices registered on one network but utilised on visited networks; and interoperability and open access to data and services, encompassing both technical and regulatory dimensions.

Personal Data Protection and Intelligent Infrastructures: The Value Contributed by Privacy Enhancement Technologies

Privacy Enhancing Technologies facilitate the integration of GDPR principles inside intelligent infrastructure systems and Internet of Things (IoT) devices. This section provides a concise

overview of contemporary Privacy-Enhancing Technologies (PETs) classified into six categories: digital signatures, authentication, communication systems, encryption technologies, computations, and general anonymisation technologies. We delineate the compliance of each PET area and methodology with GDPR processing guidelines.

Analysis of Contemporary Privacy Enhancing Technologies That Facilitate Gdpr Compliance

Digital Signatures that Enhance Privacy. Privacy-preserving digital signatures mostly adhere to the GDPR processing standards outlined above, including minimisation, integrity, and partially secrecy. Privacy-enhancing digital signatures enable users to sign communications while ensuring integrity, validity, and non-repudiation, akin to conventional digital signatures, but also offering supplementary privacy characteristics. For example, group and ring digital signatures provide signer anonymity or signer pseudonymity. Any user, as a member of the group, may thereafter sign a message anonymously on behalf of the group. Group and ring signature systems are often used in group-based authentication contexts to guarantee data integrity and validity while preserving the anonymity of signers. The legitimate signatures must be untraceable to a specific individual using the secret key. Signatures may be authenticated by any individual using a single public group key that does not identify a specific signer. Ring signatures resemble group signatures but are based on decentralised models devoid of a group manager. Group and ring signatures serve as effective foundational cryptographic mechanisms for systems necessitating user anonymisation, including e-voting, e-payments, e-coins, and many privacy-preserving applications. Certain group and ring signature techniques are included in the standard ISO/IEC 20008-2:201369.

Alternative privacy-enhancing digital signatures include blind digital signature techniques that facilitate the concealment of the message content. Blind signatures are used in scenarios when the message proprietor and the signer are distinct entities. The signer is often a third party that must not have access to the data content. The signatures are then verified by the public against the unblinded message, functioning as a conventional digital signature. Blind signatures are mostly used in payment systems like PayCash.

Authentication that enhances privacy. Privacy-enhancing authentication solutions, including Attribute-Based Credentials (ABC) and anonymous credentials, as well as anonymous and pseudonymous entity authentication protocols, mostly adhere to the data minimisation principle and partially to the purpose restriction concept of GDPR processing. These authentication methods often allow users to access services without disclosing their true identities and personal information. Certain systems additionally provide unlinkability and untraceability to avert the profiling of users' activities inside a service. The aforementioned standards and concepts are delineated in the standard ISO/IEC 29191:201272. Nonetheless, anonymous authentication solutions must guarantee the revocation of fraudulent users from the system. For example, ABC techniques rely on personal attributes rather than user identity (i.e. full name, unique identifier, digital certificate X.509). Digital identity is defined as a collection of qualities (personal attributes) that delineate an individual, including elements such as a driver's license, age, and group membership, among others. These features may be shown selectively, anonymously, and without the possibility of tracing or linking the displayed transactions. For instance, individuals seeking entry to liquor establishments must demonstrate compliance with age restrictions by providing

evidence of their age (i.e., over 21 years old). The qualities are often issued by a third-party trusted entity or service provider. ABC schemes and anonymous credentials represent typically based on asymmetric cryptographic primitives and adhere to a user-centric methodology.

Systems for Enhancing Privacy in Communication. Privacy-enhancing communication methods and systems primarily aim to ensure integrity and confidentiality during data transmission over communication networks. Common security protocols, like IP Security (IPsec), Transport Layer Security (TLS), and Secure Shell (SSH), provide verified encryption for data in client/server or peer-to-peer communications, therefore preventing eavesdropping on sensitive and personal information. However, these protocols often need the authentication and identification of data senders and recipients. Mix networks, proxies, and onion routing methods facilitate the establishment of anonymous communication networks that safeguard against intricate traffic analysis. Senders may convey messages to recipients without disclosing their identify or location. This aids in the reduction of the GDPR principle. For instance, mix networks include mix nodes (proxy servers, relays) that aggregate messages from several transmitters to obfuscate the correlation between incoming and outgoing data. Onion routing uses an onion encryption methodology whereby a sender creates a distinct encryption layer with each network node along the route, referred to as an onion router. The communications are enveloped by the sender in many layers of encryption, much to the layers of an onion. Each onion router along the route decrypts its own onion layer and forwards the data to the subsequent onion router. Upon decryption of the last layer, the data arrives at its destination (e.g., web server). Mixnets and onion routing primarily provide anonymous communication for persons on the Internet.

Technologies for Enhancing Privacy via Encryption. Encryption is seen as a security safeguard under the GDPR and a means for organisations to adhere to data protection mandates and mitigate fines in the event of data breaches. Nonetheless, the GDPR does not impose specific encryption requirements.

Privacy-enhancing encryption technologies adhere to the principles of privacy by design and by default, ensuring that stored data remains encrypted and providing superior protection assurances. Attribute-based encryption, homomorphic encryption, and searchable encryption are sophisticated cryptographic systems that provide additional features. Each of these technologies may provide outsourced data storage in an encrypted format. Attribute-based encryption provides sophisticated access control capabilities, since the encrypted data can only be decoded by an entity possessing a designated set of characteristics. Homomorphic encryption techniques provide calculations on encrypted data, ensuring that both the stored data and the results of the computations remain unavailable to the hosting service. Searchable encryption systems provide the execution of search queries and the extraction of statistics without the need of decrypting the material. Furthermore, it may be used to identify data items pertinent to certain characteristics and objectives. Authorised parties may then decrypt these data items, facilitating restricted access just to the relevant elements within the dataset. Searchable encryption and homomorphic encryption provide user-centric methodologies, allowing the data owner to retain control over the outsourcing and administration of the encrypted information.

Although encrypted datasets are inherently safeguarded against unauthorised access, data minimisation must still be implemented when selecting data items for inclusion in the encrypted dataset, as segments of the data may be decrypted, potentially allowing for re-identification if the decrypted information contains identifiable details. The encryption of a dataset does not eliminate the need to refrain from collecting and documenting any superfluous identifying information. Regarding security, encrypted data stays unattainable for the storage provider and is simultaneously safeguarded against information leaks in the event of a data breach. Data anonymity may be preserved, provided that anonymous authentication methods are used for accessing the storage service. The secrecy of the encrypted data is safeguarded; nevertheless, supplementary security methods, like as digital signatures, must be used to ensure the integrity and validity of the material.

Privacy-Enhancing Computations. Privacy-enhancing computations may be accomplished via secure multi-party calculations, which allow many parties to collaboratively calculate a function on their inputs while maintaining the confidentiality of those inputs, without necessitating the involvement of a Trusted Third Party in the computation. In a secure multi-party computing protocol, no participant may acquire information beyond their own input, the public function used for calculation, and the outcome of the collective computation. Consequently, secure multi-party calculations primarily adhere to the GDPR concept of minimisation.

Secure multi-party computing protocols are seen realistic and relevant to modern issues, such as electronic voting and auctions, owing to the significant improvement in efficiency achieved over the last decade. A standard multi-party calculation protocol ensures input confidentiality, meaning that no information about the private data of the participating parties can be deduced. The accuracy of the output may be assured, contingent upon the security model of the protocol, either by guaranteed proper output or by terminating the calculation in the event of an error 80.

Universal Anonymisation Technologies. Both classic statistical disclosure control (SDC) procedures and newer differential privacy (DP) technologies may be considered measures used by data controllers to meet compliance requirements of data protection, such as those outlined in the GDPR. Initially, we outline the workflow of these mechanisms and approaches, subsequently identifying their potential to facilitate compliance with the processing principles of the GDPR regarding personal data protection.

It is generally assumed that a data controller will maintain a database including users' unencrypted data for both SDC and DP. The data controller may anonymise the data using several techniques, such as suppressing and generalising characteristics for statistical disclosure control and calibrating noise in the query results. A key distinction between SDC and DP is that SDC allows a data controller to disseminate an anonymised database, while DP emphasises the postprocessing of query results derived from the plaintext database, without publishing any database.

SDC methods may aid in adhering to the purpose limitation principle by appropriately processing the relevant properties. If an attribute is irrelevant to the specified goal of data processing, it may be omitted prior to the publication of the database. Likewise, DP may provide contributions in two ways. It may be programmed to respond just to enquiries that align with the predetermined objective. Conversely, it may provide a meticulous implementation of purpose restriction, for instance, by regulating the quantity of calibrated noise in relation to the particular objective of data

processing. SDC mechanisms may partially mitigate data leakage by concealing critical features inside user records. Nonetheless, SDC may be vulnerable to linkage attacks or similar threats if the attacker had certain prior knowledge. Conversely, DP techniques may effectively minimise data exposure by modifying the calibrated noise, while avoiding the weaknesses associated with SDC procedures. Implementing SDC and/or DP allows a data controller to operate in a responsible manner while handling sensitive personal data. Specifically, using Data Protection, a data controller may achieve a high level of confidence irrespective of the attacker's expertise.

SDC and DP do not explicitly tackle the issues of storage limits, transparency, integrity, and secrecy. Specifically, they do not provide a typical secrecy guarantee for the data (e.g., those derived from encryption methods). Nevertheless, they provide specific degrees of privacy safeguards for the relevant human users. The accuracy principle is not strongly related to SDC and DP. Nonetheless, it is important to acknowledge that if a data controller has disseminated an anonymised database derived from SDC, it may be required to update the public database to comply with this principle. It is noteworthy that the publication of a (nearly) identical database on two occasions may pose privacy problems. The data controller must use caution if this transpires.

We assume that the data controller has access to the plaintext database for both SDC and DP. In reality, consumers may be hesitant to exchange data directly with the data controller. Furthermore, the data controller may mitigate its duty in safeguarding users' data. Consequently, distributed versions of SDC and DP may be used. Accomplishing this affords users enhanced protection.

Illustration of Privacy Policies: A Linguistic Technique for Editing, Analysis, and Enforcement of Privacy Criteria

The acquisition, use, and dissemination of user data are often governed by privacy rules articulated in plain language, delineating particular acts that are permitted, mandated, or prohibited under certain contextual conditions. This exemplifies the regulations that each social networking service provider disseminates on its data processing sites 87.

Nonetheless, as discussed in the preceding sections, the GDPR aims to govern all forms of electronic transactions involving personal data. A scenario that adheres to the stipulations of GDPR – the generation of parking tickets – has been outlined via business processes in the preceding section. Illustrations of regulations pertaining to the record of processing (Art. 30) and the permission of the data subject (Art. 7) may be articulated in everyday language using the following expressions:

Processing Record: "the Data Controller Must Document the Parking Reservation Processing."

Data Subject approval: "In the absence of consent from the data subject for the storage of parking reservations, the data recipient is prohibited from issuing the parking permit."

While natural language (NL) facilitates end users in comprehending the permitted, obligatory, or prohibited actions regarding their data, a significant challenge arises from the inherent inability of NLs to be machine-readable, rendering automatic oversight of data usage and processing by operating entities impractical. Specifically, natural languages cannot serve as the input language for a policy-based software architecture intended for policy administration. Automated policy

analysis, which ensures the absence of conflicting data policies, and policy enforcement, the implementation of data policies during data access requests, need inputs in a machine-readable format, such as the de facto standard XACML.

To further the automated management and enforcement of privacy regulations, we provide a language-based methodology that utilises machine-oriented, English-based Controlled Natural Languages (CNLs).

Controlled natural languages (CNLs) are a category of natural languages specially designed to facilitate machine processing. A Controlled Natural Language (CNL) is fundamentally a refined language derived from natural language, characterised by a more limited vocabulary, syntax, and semantics, but preserving many of its inherent qualities. Controlled Natural Languages (CNLs) exhibit a more artificial representation for grammar and vocabulary, hence diminishing the ambiguity and complexity inherent in comprehensive languages such as English, Spanish, French, Swedish, and Mandarin.

Conceptual Network Languages (CNLs) have shown efficacy in alleviating issues associated with verbal ambiguity, since they can be readily converted into formal languages such as first-order logic or various iterations of description logic, mostly in an automated and predictable manner. A branch of CNLs designed for articulating data privacy legislation is inherently formal, originating with an established formal syntax and semantics. These languages may effectively articulate the kind of information found in software specifications, formal ontologies, corporate rules, and legal and medical restrictions.

The proposed CNL aims to diminish the obstacles to adopting legal contracts governing data sharing, typically articulated in natural language, concerning privacy assurances, while also facilitating the mapping of contracts to formal languages that enable automatic verification of the agreement and enforceable languages to allow for its enforcement. A data sharing contract, or data sharing agreement, is fundamentally a policy established between two or more companies that delineates the terms and circumstances around data sharing and utilisation.

CNL4DSA. The CNL4DSA (Controlled Natural Language for Data Sharing Agreement) facilitates the enforcement of privacy and security in electronic data sharing. CNL4DSA facilitates straightforward, but formal, delineations of many categories of privacy regulations, as enumerated below:

• authorizations, denoting the consent for subjects to execute actions on objects (e.g., on user data), contingent upon certain contextual factors;

• Prohibitions denote the restriction of a subject executing actions on an item under certain contextual circumstances;

• Obligations delineate that persons are compelled to execute acts on objects, contingent upon certain contextual situations.

CNL4DSA fundamentally enables the formal specification of the statement "subject s performs action a on object o."

Incorporating the can/must/cannot constructions into the fundamental elements of the language enables the articulation of authorisations, responsibilities, and prohibitions. The authorisations, duties, and prohibitions may thereafter be assessed based on the characteristics of subjects and objects, such as users' roles, data classifications, temporal factors, and geographical contexts.

The two regulations concerning the parking reservation scenario, articulated before in plain language, are represented in CNL4DSA as follows:

Processing Record: "IF subject1 possesses the role of ParkingServiceProvider AND object1 belongs to the category of ParkingReservation THEN subject1 IS REQUIRED to document object1"

Data Subject Consent: "IF subject1 possesses the role of UserDevice AND subject2 possesses the role of ParkingServiceProvider AND subject3 possesses the role of ParkingLotTerminal AND object1 belongs to the category of ParkingPermit and object1 is associated with subject1 AND subject1 grants NoConsent THEN subject3 is PROHIBITED from issuing object1"

Toolkit based on CNL4DSA. CNL4DSA, first developed to articulate data sharing regulations, has shown its applicability in delineating many additional needs, including specifications for software product lines. The language is not domain-specific, since it lacks a predetermined vocabulary. Consequently, it may be used in many applications, including social networking, e-health, and emergency management situations.

The advantage of this language is in the development of several tools throughout the years, each designed for a certain function within the rule's life cycle. We will delineate each of these tools and the responsibilities included by CNL4DSA.

A textual rule, articulated in either CNL4DSA or normal language, is administered using a CNL4DSA-based toolset, first developed in 1997 and subsequently updated. The toolkit, originally consisting of a CNL4DSA Authoring Tool, a CNL4DSA Policy Analyser, and a CNL4DSA Mapper Tool, has recently been enhanced with the addition of the NL2CNL translation tool, which converts natural language rules into CNL4DSA rules. We provide a concise description of the components below:

• NL2CNL Translator: a user without experience in Controlled Natural Languages (CNLs) may modify rules in natural language (e.g., English); with minimum user effort, the translator generates the rules in CNL.

• CNL4DSA Authoring Tool: an author proficient in Controlled Natural Languages (CNLs) may directly modify rules inside CNL4DSA. The regulations are limited by CNL4DSA constructs (see to Section 3.7.1), and the terminology inside the regulations derives from certain vocabularies.

The CNL4DSA Analyser evaluates a collection of CNL4DSA rules, identifying potential conflicts whereby two or more rules concurrently permit and prohibit access requests to data under identical contextual circumstances. Upon detection of a disagreement, a conflict resolution approach that prioritises rules is used to accurately enforce the priority regulations.

• CNL4DSA Mapper: it converts the CNL4DSA regulations into a legally binding language. The mapping process receives the analysed CNL4DSA rules, converts them into a XACML-like

language, and integrates all the rules according to the established conflict resolution procedures. The result of this instrument is a binding policy. This policy will be assessed with each request to access, use, and manipulate the data delineated within the policy.

A CNL4DSA Lifecycle Manager coordinates all the above components. Upon logging into the Lifecycle Manager, users activate their designated functions based on their respective roles (e.g., end-data owner, data controller, data processor). Consequently, users engage with the tools via the Lifecycle Manager.

The depth and versatility of CNL4DSA, in terms of articulating requirements across many domains and possessing diverse specifications processing tools, contribute towards achieving an Comprehensive framework for the definition and study of safety, security, privacy, and trust in intricate and evolving contexts.

Conclusion

This article primarily examines the legal handling of personal data within an Intelligent Infrastructure context. The data controller is bound by many duties regarding the management of personal information. We concentrated on the stipulations of purpose restriction, data minimisation, storage limitation, integrity and confidentiality of the data, and ultimately, transparency for the data subjects.

Information and communication technology are sometimes seen as inherently invasive to privacy. Nevertheless, the GDPR assigns a significant role to ICT technologies in strengthening the principles related to personal data protection. The GDPR establishes the necessity for privacy by design. The data controller must adopt suitable technological and organisational precautions both during the determination of processing methods and during the processing itself. Privacy Enhancing Technologies are very important as they serve to unify and reinforce data protection standards across diverse ICT applications.

Privacy-enhancing digital signatures and authentication methods are known strategies that may maintain diverse privacy and security attributes for users in different ICT applications, addressing several GDPR privacy principles.

Privacy-enhancing encryption methods and privacy-enhancing computations serve as essential components for privacy-preserving apps that adhere to the principles of privacy by design and by default.

References

[1] Nguyen, D. C., Pathirana, P. N., Ding, M., & Seneviratne, A. (2020). Integration of blockchain and cloud of things: Architecture, applications and challenges. *IEEE Communications surveys & tutorials*, *22*(4), 2521-2549.

[2] Gudepu, B. K., & Jaladi, D. S. (2018a). The Role of Data Profiling in Improving Data Quality. *The Computertech*, 21-26.

[3] Bousquet, A., Briffaut, J., Caron, E., Dominguez, E. M., Franco, J., Lefray, A., ... & Uriarte, M. (2015, December). Enforcing security and assurance properties in cloud environment. In *2015*

IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC) (pp. 271-280). IEEE.

[4] Chennareddy, R. K. (2021). Designing Data and Analytics Ecosystems for High Volume Transaction Processing Applications. *International Journal of AI, BigData, Computational and Management Studies, 2*(2), 95-106.

[5] Gudepu, B. K., & Jaladi, D. S. (2018b). The Role of Data Quality Scorecards in Measuring Business Success. *The Computertech*, 29-36.

[6] Hamlen, K. W., Kagal, L., & Kantarcioglu, M. (2012). Policy Enforcement Framework for Cloud Data Management. *IEEE Data Eng. Bull.*, *35*(4), 39-45.

[7] Sethuraman, P. (2022). Latency-Aware Scheduling and Resource Control Algorithms for Emergency and Public Safety Wireless Networks. *International Journal of Emerging Research in Engineering and Technology, 3*(4), 133-140.

[8] Chennareddy, R. K. (2020). Engineering Intelligence Systems Using Big Data and Cloud Architectures for Modern Data Intensive Applications. *International Journal of AI, BigData, Computational and Management Studies, 1*(2), 41-50.

[9] Goriparthi, R.G. (2022) Interpretable Machine Learning Models for Healthcare Diagnostics: Addressing the Black-Box Problem. *Revista de Inteligencia Artificial en Medicina. 13*(1): 508-534.

[10] Gudepu, B. K., & Jaladi, D. S. (2021). GDPR Compliance Challenges and How to Overcome Them. *International Journal of Modern Computing, 4*(1), 61-71.

[11] Zhang, H., Tan, H. B. K., Zhang, L., Lin, X., Wang, X., Zhang, C., & Mei, H. (2011). Checking enforcement of integrity constraints in database applications based on code patterns. *Journal of Systems and Software*, *84*(12), 2253-2264.