Privacy-by-Design Data Engineering Framework for Multi-Cloud and Hybrid Data Platforms

(IJCSNT) International Journal of Communication Systems and Network Technologies
ISSN Number: 2053-6283
Volume 13, Issue No. 3, 185–217

https://journals.mirlabs.in/index.php/ijcsnt
DOI- 10.18486/ijcsnt/13.3.187
IJCSNT

Harshavardhan Peddireddy[1]

Abstract

Enterprises are heading into multi-cloud and hybrid cloud environments, which have enhanced capabilities for enterprise data management, but have also complicated privacy and security issues, governance models, and inter-cloud communication and regulatory compliance. The paper introduces an Enterprise Data Engineering Architecture for Privacy-by-Design that combines privacy-aware data ingestion, adaptive policy enforcement, secure data transformation, governance management, compliance verification and intelligent resource orchestration into a single framework. The framework is tested for effectiveness against the TPCx-BB (BigBench) benchmark dataset to give it a realistic test with enterprise workloads of structured, semi-structured and unstructured data. Experimental results show that the proposed framework achieves 99.11% of privacy compliance rate, 2.27 s of query execution time, 17.84 GB/s of data processing throughput, 63 ms of access latency, 331 MB of communication overhead, 99.18% of governance consistency, 99.14% of metadata synchronization accuracy, 99.08% of policy enforcement accuracy, 90.8% of resource utilization and 1.00 of scalability compliance index which exceed most of the evaluation metrics used in comparing with existing enterprise data engineering platforms. Moreover, framework shows only 0.46% privacy leakage, 99.24% unauthorized access detection, 99.73% secure data transfer, 6.18% encryption overhead, 99.58% success rate of the audit and 0.37% success rate at policy violations, thus reinforcing its capability of providing secure, privacy-preserving, and scalable enterprise data engineering for enterprise data and services in modern multi-cloud and hybrid cloud environments.

Keywords

Privacy-by-Design, Data Engineering, Multi-Cloud Computing, Hybrid Cloud, Data Governance, Privacy Preservation, Secure Data Analytics, Enterprise Data Platforms.

Received: 12 August 2024; **Revised:** 27 November 2024; **Accepted:** 12 December 2024; **Published:** 31 December 2024;

[1]Lead Solution Architect, Meijer INC, Michigan, USA.

Corresponding author:
Email: harshavardhan.mca@gmail.com

INTRODUCTION

Digital transformation has surged in responsibility across enterprise, producing an unprecedented surge in the amount, variety, and velocity of data emanating from enterprise applications, Internet of Things (IoT) devices, cloud services, and distributed information systems[1]. In order to manage these large-scale heterogeneous data efficiently, many organizations are turning to multi-cloud and hybrid cloud applications that integrate public and private cloud infrastructure for higher scalability, flexibility, fault tolerance, and cost savings. These clouds allow companies to run workloads on a mix of cloud services and private hardware to keep key business processes on premise. While these architectures enhance operation efficiency considerably, they also raise intricate issues related to the safe handling of data, preserving privacy, adherence to regulations, and uniform management in geographically dispersed cloud environments[2]. Thus, safeguarding sensitive enterprise data across the data engineering lifecycle is a non-negotiable essential for today's cloud-native apps.

Background and Motivation

Data engineering is the backbone of enterprise analytics, enabling data acquisition, integration, transformation, storage, governance, and processing to facilitate enterprise decision-making applications. Centralized, traditional clouds have the advantage of having data in fewer clusters, which makes security and privacy controls easier to implement[3]. However, this ideal has been overturned by the ubiquitous use of multi-cloud and hybrid models that spread data over disparate infrastructure and provisions set up different security policies, governance structures, and compliance standards. Information is frequently migrating between different cloud providers when being ingested, processed and analysed creating the potential for unauthorized access, privacy leakage, inconsistent policy enforcement and regulatory violation.

Compliance with the privacy regulations is a necessity for any organization across a variety of industries, including healthcare, finance, government, and manufacturing, and at the same time, they need to support data-sharing and collaborative analysis[4]. Current cloud-based data engineering approaches tend to focus on scalability and processing capabilities and often consider privacy as a secondary constitutional measure for extending security beyond the system's boundaries toward the entire data lifecycle. This constraint encourages building privacy-aware data engineering architectures based on built-in security, governance, and compliance[5]. By incorporating Privacy-by-Design, an approach allows organizations to safeguard the sensitive data they hold without sacrificing the flexibility and efficiency that come from today's multi-cloud environments.

Figure 1 shows the evolution of enterprise data platforms across traditional on-premise infrastructure, private cloud, public cloud, hybrid cloud and multi-cloud configurations. Each phase increases the scalability, flexibility, resilience, and utilisation of resources, while adding new governance and privacy needs[6]. The evolution exemplifies how the demand for data engineering solutions that are secure and privacy-preserving, and can effectively manage the growing amount of enterprise data in distributed environments over heterogeneous cloud environments, continues to rise.

Challenges in Privacy-Preserving Multi-Cloud Data Engineering

Although the benefits of multi-cloud and hybrid computing abound, it also faces certain technical hurdles in customer deployment for enterprise-scale data engineering systems. Governance can be complex,

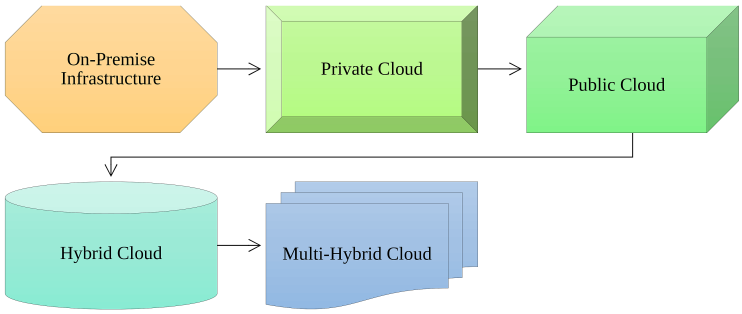

Figure 1. Evolution of Enterprise Data Platforms

as data is spread across widely varied cloud platforms, while each service may have different privacy policies. Migration of data across cloud makes the communications overhead larger and more chances of information leakage during migration and synchronization [7]. However, keeping metadata secure and maintaining data lineage through multiple platforms is also difficult as each cloud provider will offer different storage architectures and access control mechanisms.

Another challenge is implementing dynamic privacy policies that evolve based on the changing user roles, data sensitivity level, and regulations while maintaining minimal computational complexity. Conventional access control methods are static and unable to react efficiently to a changing enterprise environment. Moreover, data operations spanning distributed infrastructures need continuous monitoring, auditing and validation to preserve regulatory compliance, adding to the management challenge [8]. The need for efficient use of resources is also a point of concern; privacy preserving mechanisms cannot make too much impact on query executions performance, analytics through puts or system scalability. The challenges underscore the need for resolving privacy protection, governance, security, and computational efficiency throughout the entire data engineering pipeline in an integrated manner.

Research Contributions

The main contributions of this work are enumerated as follows:

- Multi-cloud and hybrid cloud data environment engineering framework for secure data management is being developed.
- A data engineering pipeline capable of ensuring data privacy is built by combining secure data ingestion, adaptive data classification, transformation, governance and cross-cloud processing into an integrated architecture.
- Dynamic enforcement of privacy policies, and context-based access control mechanisms are built in to enhance the protection from illegal access to the data and to enable flexible enterprise operations.
- An intelligent compliance verification mechanism is added for automatic privacy validation and governance throughout the distributed data lifecycle [9].

- The proposed framework is tested on the publicly-available TPCx-BB (BigBench) benchmark dataset and is shown to be suitable for enterprise-level analytics workloads.
- The work made through its comprehensive experimental comparisons proves improvements in privacy compliance, processing efficiency, resource utilization, scalability, and secure cross-cloud data processing against current practices.

Overall, these contributions collectively provide a practical and scalable way of enterprise data engineering in today's distributed cloud setting under privacy constraint.

The rest of this paper is structured as follows: Section 2 summarizes recent works on multi-cloud data engineering, hybrid cloud platforms, Privacy-by-Design principles and presently available privacy preserving data management techniques, presenting current research gaps. Section 3 will address the system architecture, design goals, threat model, and problem formulation for the multi-cloud data engineering system. In section 4, the proposed Privacy-by-Design Data Engineering Framework is explained with its architecture, privacy-aware processing pipeline, governance mechanisms, policy enforcement strategy, secure cross-cloud data management, and proposed algorithm. Section 5 introduces the TPCx-BB dataset, the experimental configuration, the performance evaluation measurement, the comparative analysis, the privacy evaluation, and scalability evaluation and discusses the experimental results. Finally, Section 6 finalizes the paper and introduces directions for future research for better supporting privacy-pervading enterprise data engineering in evolving multi-cloud landscapes.

RELATED WORK

Multi-cloud technology and hybrid cloud structures have revolutionized enterprise data engineering by allowing organisations to manage large amounts of structured, semi-structured and unstructured data in geographically distributed computing infrastructure[10,11]. These architectures will make systems more scalable, available, and fault-tolerant and less reliant on a single cloud vendor. However, multiple cloud systems with different platforms and configuration poses issues with regard to privacy preservation, governance uniformity, regulation compliance and secure data processing. The opportunity to implement data engineering, secure analytics, privacy protection, and metadata management within the cloud has inspired a number of recent studies that offer solutions[12]. However, current methods typically focus on only part of these goals or leave large research gaps to develop multi-faceted privacy-aware data engineering frameworks. Recently, several new developments in multi-cloud data engineering, data processing with Hybrid Cloud Data Management, Data Processing with Privacy-by-Design and governance mechanisms have been discussed that can help bridge the gap, and also examine which gaps are left untouched by this research.

Data Engineering for Multi-Cloud Platforms

To ensure operational resiliency and optimize resource use, modern mission-critical enterprise applications are becoming more multi-cloud. Multi-cloud data engineering frameworks are mainly developed to ingest data across cloud providers, store the data with scalability, distribute it in parallel, and balance workloads across cloud providers[13]. The adoption of distributed data technologies, including Apache Spark, Apache Flink, Apache Hadoop, Delta Lake, and Apache Iceberg, has revolutionized distributed data analytics, enabling large-scale data processing with fault tolerance and elastic resource provisioning.

While these platforms offer performance benefits and rewrite mainly on performance, they focus on data lifecycle management with privacy. Privacy controls are typically applied either during data ingestion or at access authorization time, leaving gaps during data storage and transformation, as well as synchronization [14]. Moreover, there are many different cloud infrastructures and these have incompatible metadata and access policies that makes unified governance hard. Thus, the integrated privacy protection that supports the data throughout all processing phases becomes essential for modern approaches to data engineering, alongside endpoint security.

Table 1. Literature Summary of Multi-Cloud and Privacy-Preserving Data Engineering Approaches

Method	Platform/Technology	Contribution	Limitation
Apache Spark	Distributed Data Analytics [15]	Enables large-scale parallel data processing with fault tolerance.	Limited built-in privacy enforcement mechanisms.
Apache Flink	Stream Processing	Supports low-latency real-time analytics and event processing.	Insufficient cross-cloud governance capabilities [16].
Apache Hadoop	Distributed Storage	Provides scalable storage and batch processing for big data workloads [17].	High resource consumption and weak privacy integration.
Delta Lake	Data Lake Architecture	Improves data consistency through ACID transactions and version control.	Does not provide adaptive privacy policy management [18].
Apache Iceberg	Metadata Management [19]	Supports schema evolution and efficient metadata handling.	Limited compliance automation for distributed clouds.
Databricks Unity Catalog	Enterprise Data Governance	Centralizes metadata management and access permissions [20].	Mainly designed for a unified cloud ecosystem.
Kubernetes-based Orchestration	Cloud Resource Management	Dynamically allocates resources across distributed cloud environments.	Privacy-aware scheduling is not considered [21].
Data Mesh Architecture	Decentralized Data Management [22]	Enables domain-oriented ownership and scalable enterprise analytics.	Governance consistency remains difficult across domains.
Data Fabric Framework	Intelligent Data Integration	Connects heterogeneous data sources through unified data services.	Privacy mechanisms are mostly external to the framework [23].
Lakehouse Architecture	Unified Analytics Platform	Combines warehouse reliability with data lake scalability [24].	Secure cross-cloud synchronization remains limited.

A compilation of key technologies used in enterprise data engineering is summarized in Table 1. While these methods give a great boost to distributed data processing, scalability, metadata management, and cloud orchestration, the majority of them currently lack privacy-by-design, adaptive governance, and automated compliance capabilities across diverse multi-cloud environments.

Hybrid Cloud Data Management

Hybrid cloud solutions integrate private cloud and public cloud to offer a mix of high performance, security and cost. Cloud resources like scalable analytics and backup services are often used in public cloud along with the use of private cloud for retention of confidential information in the cloud [25]. These deployment approaches offer flexibility but complicate data movement, data synchronization, metadata consistency and management of access.

The latest hybrid cloud management models focus on workload scheduling, smart data placement, optimization of storage, and data disaster recovery [26]. But, private and public cloud synchronization can be difficult because data often span administration lines with alternate security models. As enterprise data grows, a consistent access policy becomes harder to build, monitoring results are more complex in the case of distributed transactions and data lineage remains difficult to maintain is complete. Therefore, hybrid cloud data engineering demands that data be given a unified orchestration approach that is secure for handling a distributed environment of data assets while still providing high analytical performance.

Privacy-by-Design Principles

Finally, Privacy by Design has become a design paradigm where privacy protection is built into the system's design instead of being added to it after installation [27,28]. This approach is about proactive privacy preservation by keeping data to a minimum; having secure default settings; transparency; accountability, and continually evaluating risk. As private laws have become stricter in regards to enterprise data management, these philosophies have now become more prominent.

Current applications of Privacy-by-Design are mainly concerned with hidden stages of data processing such as encryption, anonymisation, pseudonymisation, and access control settings. These techniques of enhancing confidentiality work well in isolation but lack integration of privacy over the entirety of the engineering life cycle [29]. The implementation is even more complex in distributed cloud environments due to dynamic movement of data across multiple infrastructures and the need for adaptive policies and automated compliance checks. Hence, enterprise-level data engineering requires comprehensive Privacy-by-Design designs that enable privacy preservation at ingestion, analytics and data sharing.

Data Governance & Compliance Frameworks

Strong data governance is the assurance that enterprise information stays accurate, secure, consistent and compliant throughout its lifecycle. The elements of governance usually contain metadata, data lineage tracking, data policy enforcement, data auditing and monitoring compliance with laws and regulations [30]. Businesses, especially large businesses, are looking to implement automated governance solutions to meet today's legal demands at the same time as increasing transparency in an enterprise.

There are also a few enterprise-level governance metadata repositories and catalog services available on the market as well as open source. These systems help to provide transparency on the existing data assets spread across the distributed resources; however, they may not have adaptive privacy management features defined as needed for heterogeneous multi-cloud environments. Within many organizations, manual compliance validation is still widely practiced, adding to administration and delay of regulatory reporting [31]. Combining intelligent governance with automated privacy validation is thus vital to ease operation in order to ensure ongoing compliance within the distributed infrastructures.

Secure Data Processing Techniques

Enterprises deploying analytics across multiple cloud platforms are obliged to safeguard data processing. Current approaches are such as encryption at rest and in transit, attribute-based access control, secure authentication, tokenization, trusted key management, differential privacy, and trusted execution environments[32]. These mechanisms shield the sensitive information from unauthorized disclosure and facilitate collaboration on analytics.

Although significant advances have been made, a lot of secure processing systems add on some extra computation industries which lessen the full speed of the query and expand all round assets utilization. Moreover, privacy enforcement is often not tightly coupled with the metadata management or governance workflow, limiting the interoperability between various cloud infrastructures[33]. Adaptive security mechanisms that can automatically adjust to changing sensitivity, role and compliance remain very limited. As for future enterprise data engineering, security, governance, and intelligent orchestration should be combined into a single framework model that will ensure data privacy while adhering to analytical efficiency.

Research Gap and Motivation

The literature review reveals the following research gaps that call for the proposed framework:

- Current multi-cloud data engineering platforms focus on scalability and performance features and include only partial end-to-end privacy protection.
- The majority of hybrid cloud management plans do not support a single privacy policy across various cloud vendors.
- Current privacy-by-design solutions are typically focused on siloed security solutions rather than the entire data engineering lifecycle.
- Distributed enterprise data platforms are still not well connected to automated verification of compliance and intelligent governance.
- The existing secure data processing techniques use more computational overhead, impacting on the analytical performance and resource utilization.
- Maintaining cross-cloud metadata synchronization and data lineage management are still difficult in a heterogeneous environment.
- Existing data engineering frameworks often do not include any adaptive privacy policies that can meet dynamic enterprise requirements.
- There are few studies that assess privacy preserving data engineering on a standardized enterprise benchmark datasets representative of real-world multi-cloud workloads.
- An all-encompassing, privacy-aware ingestion, governance, and secure processing framework, identity and validation of compliance and cross-cloud orchestration is still missing within a single architecture.

The development of the proposed Privacy-by-Design Data Engineering Framework is based on these research gaps, as it is designed to support enterprise multi-cloud and hybrid data platforms with high scalability and processing efficiency while simultaneously delivering adaptive privacy preservation, intelligent governance, secure cross-cloud processing and automated compliance assessment[34].

Privacy-by-Design Data Engineering Framework

The Privacy-by-Design Data Engineering Framework defines the unified architecture to build heterogeneity-based multi-cloud and hybrid environments and introduce privacy protection across the entire data lifecycle in an enterprise. The fundamental guideline of the framework is that privacy shouldn't be added as a separate security measure once data has already been gathered or processed. Instead, privacy is addressed right at the point of data entry and then cascaded throughout the platform's various classification mechanisms, transformations, governance, storage, access control, cross-cloud processing, compliance verification, and resource allocation.

The framework is supported by enterprise data that is available from different sources, ranging from transactional databases, enterprise applications, cloud repositories, application programming interfaces, business information systems, streaming platforms, and distributed storage services. These sources might contain data of varying degrees of sensitivity and applying the same privacy control across all records would incur excessive processing overhead. However, inadequate protection could leave potentially sensitive information exposed. The framework adopts an adaptive approach, basing the level of privacy on the nature and context of specific data assets.

There are a number of logically related stages in the processing pipeline. To validate and classify information coming in, on a privacy sensitivity scale. Second, privacy preserving transformations make the data suitable for analysis while maintaining sensitivity labels and provenance data. Third, the metadata, lineage, ownership, and information on policies are maintained by a governance mechanism. Access Policies are then applied to decide who or what can execute requested operations and operations it cannot execute. Cross cloud processing defines how the information flows among the cloud environments and continuous compliance verification is checking if cloud processing activities comply with relevant policies. The orchestration of resources is the final step in choosing appropriate computing resources, taking into account privacy requirements as well as performance.

This architecture thus posits privacy, governance, security, and computation efficiency as interrelated aspects of data engineering as opposed to administrative functions.

Framework Overview

The framework could be seen as a lifecycle-based data processing paradigm where each enterprising data object is provided with its actual information and affixed privacy metadata. The ingestion mechanism does not forward information to analytical applications right away when the information arrives from an enterprise source. However, when the data are examined, they are checked for their respective validity, integrity, source attributes and sensitivity, as well as governability requirements.

The first operation is to create a coherent view of data from a variety of sources, also known as a heterogeneous enterprise data source. The information received from various data sources can vary significantly both in schema and in the format, volume, quality and sensitivity. For instance, some cloud repository might have structured transaction logs, while some other source may offer semi-structured operational logs. The framework then makes it, therefore, the object of the incoming environment into a collection of records—one by one to which it can be evaluated.

The figure 2 shows the end-to-end workflow of the proposed Privacy-by-Design Data Engineering Framework for multi-cloud/hybrid cloud platforms. Enterprise data are ingested securely, classified, transformed and governed before adaptive privacy policies control cross-cloud processing of data. This

ongoing compliance monitoring and intelligent resource usage ensures privacy, governance conformity, and effective resource utilization, empowering secure business analysis while distributed data is processed with capacity and privacy safeguards.

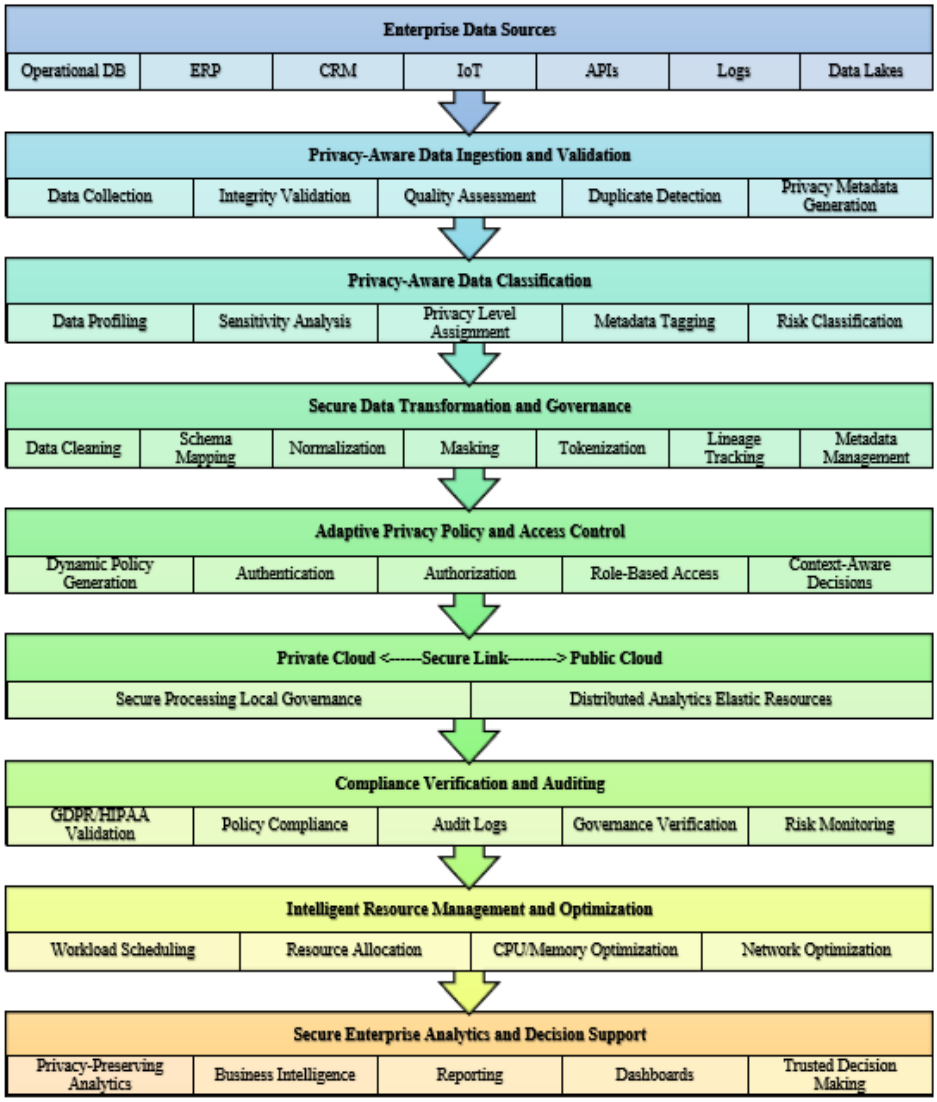

Figure 2. Overall Architecture of the Proposed Privacy-by-Design Data Engineering Framework

Let the overall enterprise data be represented as

$$D = \{d_1, d_2, \ldots, d_N\} \tag{1}$$

In which D is the full set of enterprise data entering the framework, d_i is the i^{th} data record, and N is the number of data records that are available for processing.

Each record could have several attributes and each record is represented as

$$d_i = \{a_{i1}, a_{i2}, \ldots, a_{im}\} \tag{2}$$

In which a_{ij} is the j^{th} attribute of the record d_i, and m is the number of attributes of the record di. This representation can enable privacy requirements to be assessed both at the record level and at the attribute level.

This framework carries a privacy-state, instead of considering privacy as a dichotomy between security/insecurity. The privacy state describes the current level of satisfaction of the protection requirements related to the records available. It is therefore subject to change when records are transformed, transferred, accessed, subject to further protection.

The states of aggregates are represented as

$$P_D = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{3}$$

In which N is the number of records that have been evaluated, P_i is the score assigned to the privacy protection of record d_i, and P_D denotes the privacy protection score of the dataset D. Individual P_i values can be expressed in a normalized range between 0 and 1, where 1 point towards more satisfactory fulfilment of clearly defined privacy requirements. Where P_D is another internal evaluation measure and not evidence of conforming to a specific regulation. Explicit policy constraints are used for the actual compliance, which is examined separately. Once the data representation and privacy state is determined, records are sent to the ingestion and classification phase with privacy.

Privacy-Aware Data Ingestion and Classification

Privacy protection starts at ingestion, as that is where information is first taken into the managed data engineering environment. The typical characteristics of a conventional ingestion pipeline are throughput, schema mapping, and guaranteed data delivery. The transfer of sensitive information, however, before it has been determined what type of privacy it needs can cause unnecessary exposure if placed in distributed storage. The envisioned concept thus includes privacy inspection as part of the ingestion process.

The ingestion stage consists of four main tasks: source verification, assessment/integrity check, sensitivity estimation, and privacy classification. Source verification verifies that the information in coming in is from an authorized source. Quality and Integrity Assessment identifies if the record is sufficiently complete and consistent to move forward in the pipeline. The privacy importance of information is calculated by sensitivity estimation. Last but not least, the classification functionality labels a protection category, which is then fed into transformation, governance, access-control and cross-cloud processing modules.

The quality of information received must be established before being measured for sensitivity. Poor information quality may lead to misclassification of information. Sensitive information, for instance,

might be given an incorrect level of protection because its metadata about data ownership is missing. It therefore builds up a score of ingestion validation that consists of completeness, schema consistency and source reliability.

The validation score for record di is extended by the following principle:

$$V_i = w_c C_i + w_s S_i^c + w_r R_i \tag{4}$$

subject to $w_c + w_s + w_r = 1$.

In which V_i is the score of record d_i obtained from the validation, C_i is the completeness, S_i^c is the schema consistency, R_i is the source / transmission reliability, and w_c, w_s and w_r are non-negative weights of importance of each type. Each component score can be scaled to $[0, 1]$.

Placing a record in a quarantine rather than silently propagating it into a subsequent analytical stage of records which do not meet minimum validation requirements.

Sensitivity Assessment

A processed record is then evaluated to check the privacy protection that is required. Privacy sensitivity isn't defined by whether there is personally identifiable information (PII) in a field; there are many factors to consider. The confidentiality, ownership, contractual restrictions, business significance, and/or regulatory implications of enterprise data can all contribute to the data becoming sensitive.

As a result, the framework considers several factors when calculating sensitivity:

$$S_i = \alpha Q_i + \beta O_i + \gamma R_i^g + \delta B_i \tag{5}$$

With $\alpha + \beta + \gamma + \delta = 1$.

In which S_i is the sensitivity score of record d_i, Q_i is the sensitive content score detected, O_i is ownership sensitivity score, R_i^g represents regulatory relevance, B_i business confidentiality score, $\alpha,\ \beta,\ \gamma,\ \delta$ determine the proportion of influence of the four factors.

This way, they won't be classified based on one attribute. High-level protection can be given to an enterprise record where the direct personal sensitivity is low, but there is high contractual confidentiality.

Privacy Classification

The sensitivity score is then transformed into a privacy class that can be interpreted. To achieve these three levels, a reasonable balance between policy granularity versus operational complexity is considered:

$$L_i = \begin{cases} \text{High}, & S_i \geq \tau_H, \\ \text{Medium}, & \tau_L \leq S_i < \tau_H, \\ \text{Low}, & S_i < \tau_L. \end{cases} \tag{6}$$

In which L_i represents the privacy class assigned to record d_i, while τ_H and τ_L represent the high- and low-sensitivity thresholds, respectively, where $\tau_H > \tau_L$.

Higher classification may lead to more serious transformation and access/ placement restrictions. The possibility of selective protection can be afforded to medium-sensitivity information, while low-sensitivity information can be processed with fewer restrictions. The thresholds are parameters that should be configured by organizational policy, and cannot be simply taken as given.

Redundancy and Duplicate Control

Another aspect of privacy-aware ingestion is that of unneeded data duplication. Having multiple copies of valuable files in multiple clouds also means having more places to monitor and shield. Duplication detection is thus done prior to distributed storage.

If there are two records, the similarity of these records in the attribute-set can be scaled by the Jaccard coefficient:

$$J(d_i, d_j) = \frac{\mid A_i \cap A_j \mid}{\mid A_i \cup A_j \mid} \tag{7}$$

In which $J(d_i, d_j)$ means the similarity between d_i and d_j, and A_i, A_j are the respective attribute-value sets. The lower the number, the less similarity there will be; a value near 1 signifies a great deal of overlap. In practice, it should only be used to duplicate detection attributes, not arbitrary columns that are considered sensitive information.

Once classified each record is accompanied by a privacy descriptor:

$$PM_i = \{ID_i, L_i, \Pi_i, \Omega_i, \Lambda_i, t_i\} \tag{8}$$

In which PM_i represents privacy metadata of record d_i, ID_i its identifier, L_i is the sensitivity level associated with d_i, Π_i is the applicable privacy policy identifier, Ω_i is owner information, Λ_i is lineage information, and t_i is the classification timestamp.

The key element of this step is not simply classification of data but its significant value in this stage. Establishes a long-lasting privacy context that follows data to the next Engineering work.

Secure Data Transformation and Governance

Once enterprise information is ingested and classified, it has to be converted to a format more suitable for distributed storage and analysis. In reality data is always changing—when data is extracted from various systems, especially, it often has conflicting schemas, inconsistent units, missing data, redundant fields, odd data formats and data attributes are unfamiliar to downstream applications.

But privacy can be hurt without intention, by using conventional transformation pipelines. Sensitive attributes can be copied into intermediate tables, lineage can be lost as data is aggregated, or transformed data can be given permissions that are different from the original sensitivity. The proposed transformation mechanism can then be used to preserve the privacy metadata that is created during ingestion and to conduct the necessary data preparation operations.

The transformation stage brings in data quality's correction, harmonizing data schemas, normalizing the data, privacy treatment and propagating the lineage. The treatment for privacy is chosen based on sensitivity category. Transformations may consist of suppression, generalization, masking, tokenization, pseudonymization or encryption depending on the application and the allowed analysis. Not every technique needs to be applied to each record; the technique to be supported in the selected transformation will be determined later by using the policy conditions.

Min-Max transformation, for numerical attributes that need to be normalized, can be written as

$$x_{ij}^{'} = \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_j^{\min}} \tag{9}$$

In which x_{ij} is the absolute value of numerical attribute j in record i, $x_{ij}^{'}$ is the normalized value of attribute j in record i, $x_j^{\min}$ is the minimum value of attribute j, and $x_j^{\max}$ is the Maximum value of attribute j.

Normalization is not a Privacy mechanism. It is added because the engineering pipeline needs to process the dataset in such a way that it can be used for downstream processing, without compromising the privacy controls. For example, identifiers should not be normalized numerically, but should be protected separately from normalization.

Privacy-Preserving Governance

To enable Privacy-by-Design operation, one must not rely on the sole function of transformation. The framework should keep track of source of data, transformation performed, who holds ownership, policies, and storage of transformed copies. The governance layer therefore works all the time in parallel with transformation pipeline.

The lineage repository is updated with the source object, transformation operation, execution time, processing service that is responsible, resulting data object, and destination environment for every transformation event. Thus, transformation does not sever the link between the information that is original and the information that is derived.

There are four complementary aspects to governance quality:

$$G = w_l L_c + w_m M_c + w_a A_c + w_p P_c \tag{10}$$

subject to $w_l + w_m + w_a + w_p = 1$.

In which G is the governance quality score, L_c is the lineage completeness score, M_c is the metadata consistency score, A_c is the access-policy consistency score, P_c is the policy/compliance metadata completeness score, and w_l, w_m, w_a, w_p are all non-negative weighting coefficients.

This formulation is better than a mere arithmetic average because various organizations can have different levels of significance for lineage, metadata quality, access consistency, and compliance information.

The governance mechanism thus establishes continuity with the subsequent control of the policy. The framework will not presume that converted data is insensitive if a sensitive dataset is actually converted into several derived datasets. Transformation rules are computed for labels and lineage relationships to propagate down the stream to downstream systems, and to preserve knowledge about the lineage and original protection requirements.

In a multi-cloud environment, metadata sync is also a must. The same record that is processed by one cloud can then be consumed by another cloud-based service. If this governance information doesn't sync with the destination platform, it may follow inconsistent policies regarding privacy. The governance state is thus logically maintained and can serve as a reference before decisions are made on data placement, access and transfer.

Adaptive Privacy Policy and Access Control

Once the data is ingested, classified, transformed and governed in a privacy-preserving manner, the architecture should define accessibility and usage of the enterprise's data in heterogeneous cloud environments. In a traditional cloud system, access policies are static and have a rigid structure of user and service permissions. These mechanisms offer some level of protection, but are not sufficiently effective

for multi-cloud environments as data is constantly shifting between cloud providers, organizational units, and analytical services. Security clearances are one thing from one cloud to another and other things can be introduced into the data after it has been moved or transformed. Therefore, the resulting static security policies give too much or too little access/performance.

The proposed framework adopts an adaptive privacy policy scheme, which dynamically assesses the context of each request for access and grants data access accordingly. The decision process takes into account the size of the resources, the purpose of processing, the legislative obligations already in place, the role of the user, the legislative sensitivity, and the location of the resources. These parameters can be dynamically altered during enterprise operations, and privacy policies are recomputed on each individual access of data, rather than fixed permissions being assigned before use. Through this adaptive functionality, unauthorized users can be held back, but still efficient analyses can be carried out.

An integrated score of privacy policy is calculated for each request, which combines several contextual elements.

$$PP_i = \alpha L_i + \beta U_i + \gamma C_i + \delta E_i \tag{11}$$

In which PP_i denotes the privacy policy score associated with the i^{th}-request, L_i represents the privacy sensitivity level as determined in the classification process, U_i represents the user's trust level as obtained during the classification process, C_i denotes the privacy compliance priority associated with the requested data set, E_i denotes the security level of the execution environment, and $\alpha, \beta, \gamma, \delta$ are weighting coefficients that sum to 1.

The policy score is then used to compare to organizational thresholds and determine if it is safe to operate or not. The higher the results the stronger the privacy protection, while the lower results are the more flexible the analysis of non-sensitive information can be.

The process of authorisation is represented as

$$A_i = \begin{cases} 1, & PP_i \geq T_p \\ 0, & PP_i < T_p \end{cases} \tag{12}$$

In which A_i is the authorization decision for request i, T_p is the minimum policy threshold, and 1 and 0 means permission granted and permission denied, respectively.

The policy is followed by a framework which specifies the level of access following authorisation. The proposed system uses an adaptive privilege allocation mechanism which takes into account both the sensitivity of the dataset and the characteristics of the user.

The privilege level of the access is represented by

$$PL_i = \frac{R_i + T_i + P_i}{3} \tag{13}$$

In which PL_i stands for privilege level assigned to request i, R_i is an organizational role score assigned to the user, T_i is a score on the users' perceived trustworthiness, and P_i is a score of the requestors' processing purpose as defined by enterprise policies.

Access requests are not only reviewed at first login, but are monitored throughout execution. Whenever the privacy classification and/or the governance state changes during the processing, the framework

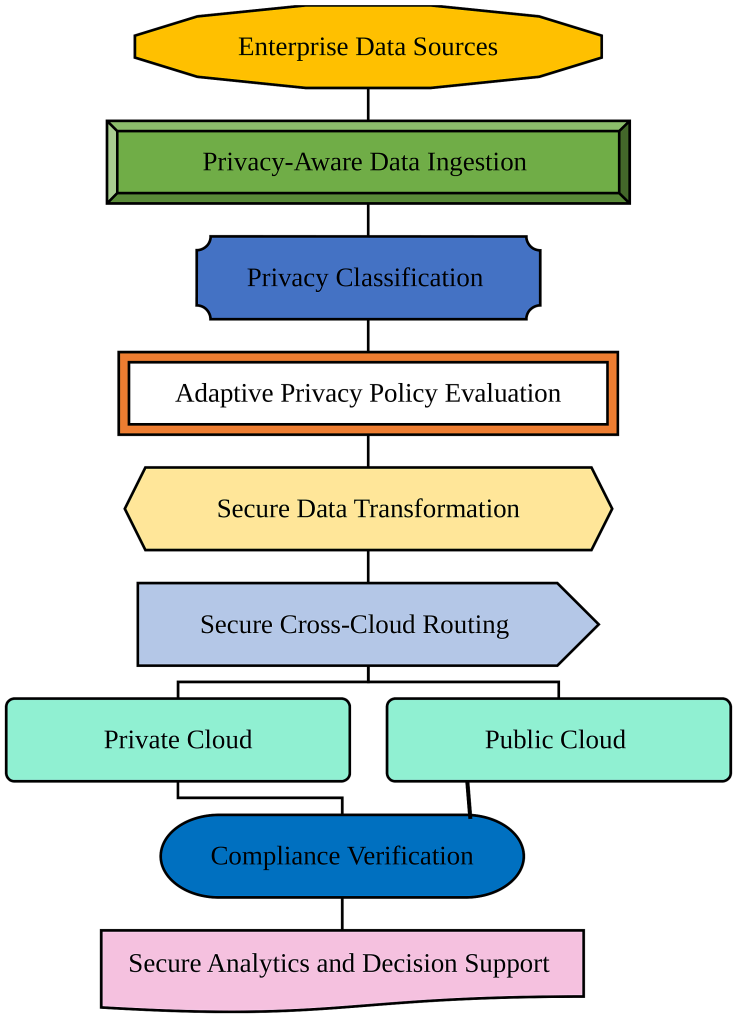

Figure 3. Workflow of Privacy-Aware Data Processing Across Multi-Cloud Platforms

will automatically recalculate the authorization decision, thus aiding continuous privacy enforcement in dynamic cloud environments.

The figure 3 shows a sequential representation of the privacy-aware data processing in a multi-cloud environment. Enterprise data are encrypted upon ingestion, classified based on privacy requirements, tested for adaptable policies, and securely communicated and transferred to a private cloud or a public cloud. Growing regulatory demands are met alongside compliance-driven value from resource usage, cloud-based security and collaboration, and trusted business decisions.

Secure Cross-Cloud Data Processing

Enterprise datasets typically span several cloud regions to enable distributed storage, collaborative analytics, backup and disaster recovery, and applications workload balancing. While cross-cloud communication is used to enhance scalability and availability, it also heightens risks of privacy leakage as sensitive information is relayed between different infrastructures that vary in security levels. The proposed framework, accordingly, refers to secure cross-cloud data process, where privacy policies are attached to the data between the clouds where data is transmitted or processed, and subsequently stored, as opposed to only where it originates.

The framework assesses the communication security of the cloud to which it receives the data before transferring it. Before entering into data movement, factors like security history, governance compatibility, network trust, authentication reliability, and encryption ability are taken into account. Thus, only when they both meet enterprise privacy requirements, the information can be transferred.

The secure transfer self-assurance among cloud environments is calculated as

$$SC_{ij} = \frac{E_{ij} + N_{ij} + G_{ij}}{3} \tag{14}$$

In which SC_{ij} is the secure communication confidence that exists between source cloud i and destination cloud j, E_{ij} is the amount of encryption capability between the two clouds, N_{ij} is the network security reliability between the two clouds, and G_{ij} is the level of compatibility existing in the governance position.

Data transfers are not made if the communication confidence does not meet the protection level requirements. Otherwise, other privacy protection methods like data encryption or policy changes will be used before migration.

The secure transfer decision is represented by

$$T_{ij} = \begin{cases} 1, & SC_{ij} \geq T_s \\ 0, & SC_{ij} < T_s \end{cases} \tag{15}$$

In which T_{ij} is the decision to transfer between cloud i and cloud j, and T_s is the minimum-security threshold for communication.

After the secure transmission, the framework calculates the best position of enterprise datasets. Information distribution is not random - taking privacy classification and computing resources into account. Data in private cloud environments may include highly sensitive information and data in lower-sensitivity analytical applications can be analysed in the public cloud.

The score for the placement suitability is given by

$$PS_i = \lambda S_i + \mu R_i + \nu C_i \tag{16}$$

In which PS_i denotes the placement suitability score of i-session, S_i the privacy sensitivity score of i-session, R_i the available computational resources at the destination, C_i involved communication efficiency score of i-session and λ, μ, ν are weighing coefficients fulfilling the condition $\lambda + \mu + \nu = 1$.

In distributed analytical processing, the processing load should be divided among the multiple cloud environments without breaking privacy restrictions. The proposed framework consequently

proposes to measure the processing efficiency by taking into account both computational potential and communication delay.

The processing efficiency is equal to

$$PE_i = \frac{Q_i}{L_i + \varepsilon} \tag{17}$$

In which PE_i is the processing efficiency, Q_i is the computational throughput, L_i is the communication latency and ε is a small positive number to prevent division by zero.

Finally, the framework assesses the security level of the overall distributed processing security, operational security indicator, comprising authorization accuracy, communication protection, governance consistency and policy compliance.

$$OS = \frac{A_c + S_c + G_c + C_c}{4} \tag{18}$$

In which OS means the total security provided in the processing, A_c means the security provided in the authorized process, S_c means the security provided in the communication process, G_c means the consistency of governance across participating clouds and C_c means the consistency of compliance as processing is distributed across clouds.

The adaptive mechanism and its ability to quicken SD processing all while providing secure cross-cloud processing create privacy protection more than just standalone checks. Access permissions change dynamically based on the continuously changing needs of an enterprise, and the privacy-preserving placement and transmission rights of cross-cloud communication stay secure. As a result, enterprise data can be processed shareable in distributed fashion across several different cloud infrastructures without the need to compromise confidentiality, control and integrity of governance, computational efficiency etc.

The figure 4 shows the intelligent governance process which continually and constantly controls multi-cloud privacy, compliance and resource optimization. Adaptive policy enforcement with metadata sync and compliance checks before allocating workloads from private to public cloud. Continuous monitoring in turn, provides feedback that brings forth dynamic adaptation of the privacy policies and resource scheduling, guaranteeing business agility and compliance with regulations, resource optimization, safe data handling, and enterprise analytics scalability.

Compliance Verification and Intelligent Resource Management

Multi-cloud and hybrid cloud enterprise data engineering systems must adhere to the governing policy of the organization and comply with regulatory requirements across the entire data lifecycle. Previous modules cover securing the data during ingest, transformation and processing but there is a need for ongoing verification to ensure that all processing operations continue to be compliant with any relevant privacy policies. Typical cloud services work with compliance audits on a periodic basis, which means that once processing is done, any violation of the compliance policy will be found. This delayed verification raises the likelihood of any data being used without authorisation and/or non-compliance with any regulations. Hence, the proposed framework features a continuous compliance verification mechanism that runs along with all phases of enterprise data engineering.

The compliance engine tracks the updates to metadata, access requests, transformation operations, cross-cloud transfers and the governance records. All processing is compared to pre-determined

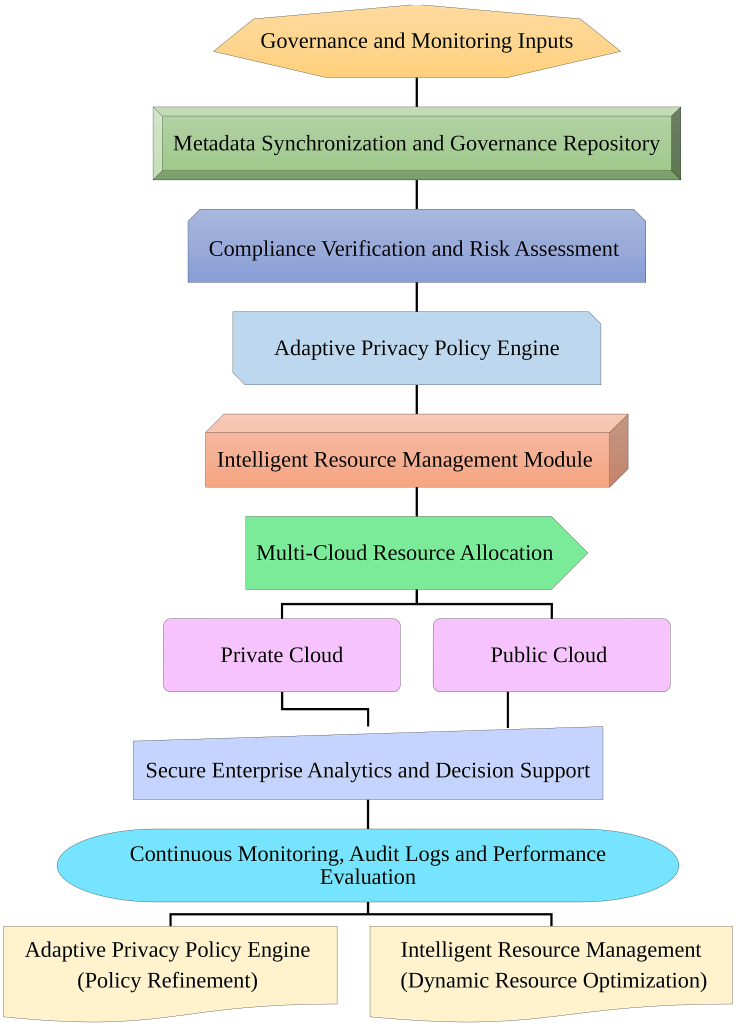

Figure 4. Intelligent Privacy Governance and Resource Management Workflow

enterprise policies before execution is completed. If it is identified that a violation is occurring, the framework could immediately shut down the operation, provide for changes in privacy policies, or move operations to another cloud supplier. This ongoing monitoring helps to lower security risks, and also lessens the workload of manual auditing.

A compliance score is determined for each processing operation by together analyzing policy satisfaction, consistency of governance, the protection of privacy, and completeness of the audit.

$$CV_i = \frac{P_i + G_i + S_i + A_i}{4} \tag{19}$$

In which CV_i stands for compliance verification score of operation I, P_i signifies policy satisfaction, G_i signifies posture in governance consistency, S_i signifies level of protection of privacy and A_i signifies audit completeness.

The calculated score is used to decide if processing meets enterprise needs.

$$C_i = \begin{cases} 1, & CV_i \geq T_c \\ 0, & CV_i < T_c \end{cases} \tag{20}$$

In which C_i is equal to the compliance decision if the operation is i, T_c is the threshold below which the compliance decision must not drop. A value of one means the compliance has been successfully verified, a value of zero means that the corrective action needs to be taken before proceeding with the process.

Once compliance verification is done, the framework then carries out an intelligent resource management. Unlike conventional cloud schedulers, which assign computational resources only based on processor availability and workload size, the proposed framework takes into consideration both privacy requirements and communication overhead as well as utilization and processing demand. As a result, workloads that demand a secure environment are processed in secure computing environments and less sensitive analytics on elastic public cloud computing.

A resource suitability score is calculated as:

$$RS_i = \alpha U_i + \beta P_i + \gamma M_i + \delta N_i \tag{21}$$

In which RS_i is the suitability score of the selected computing resource, U_i is the utilization of the processor, P_i is the privacy compatibility, M_i is the memory resources, N_i is network performance, respectively, $\alpha, \beta, \gamma, \delta$ are weighting scores corresponding to them and must satisfy the condition $\alpha + \beta + \gamma + \delta = 1$.

The distributed execution is optimised continuously to best fit the computational capacity.

$$W_i = \frac{R_i}{\sum_{k=1}^{n} R_k} \tag{22}$$

In which W_i is the work load assigned to resource i, R_i is the computational capacity of resource i, and n is the number of available computing resources.

Lastly, after the resources are allocated, the overall operational efficiency is assessed.

$$OE = \eta PE + \theta CV + \kappa RU \tag{23}$$

In which OE is operational efficiency, PE is processing efficiency provided by Equation (17), CV is compliance performance, RU is resource utilization efficiency, and $\eta + \theta + \kappa = 1$.

These operations allow the framework to dynamically optimize the stages of preserving privacy, providing governance, computational performance and resource utilization across distributed multi-cloud environments and in alignment with the regulations and laws.

Algorithm of the Proposed Framework

The complete Privacy-by-Design Data Engineering Framework is summarized in Algorithm. The algorithm keeps harmonizing privacy-aware data acquisition, secure data transformation, governance data synchronization, adaptive policy enforcement, distributed cloud processing, compliance verification, and intelligent data resource optimization in a seamless enterprise data engineering process. These operations are not performed individually, instead, they share the metadata and privacy data with subsequent operations, maintaining privacy protection in a complete processing lifecycle.

Input:

- Enterprise dataset D
- Enterprise privacy policies PP
- Governance metadata GM
- The collection of cloud resources that can be accessed CR.

Output:

- Secure processed dataset SD
- Compliance verification report (CVR)

Begin

Phase I: Privacy-Aware Data Acquisition

- Gather enterprise data from various and disparate sources
- Ensure data quality and integrity.
- Estimate privacy sensitivity
- Classify enterprise records
- Generate privacy metadata

Phase II: Secure Data Transformation

- Normalise and standardise datasets.
- Remove duplicate records
- Update metadata repository
- Maintain lineage information

Phase III: Adaptive Privacy Management

- Generate dynamic privacy policy
- Evaluate access request
- Set an adaptive privilege level
- Allow or deny access.

Phase IV: Secure Cross-Cloud Processing

- Assess the security of cloud destinations

- Choose the cloud environment best suited to the task.
- Carry out secure data migration
- Perform analytical processing using distributed processing.
- Synchronize governance information

Phase V: Compliance Verification

- Monitor policy compliance
- Validate governance consistency
- Generate audit records
- Update compliance repository

Phase VI: Intelligent Resource Optimization

- Assess cloud resources that are available.
- Allocate computational workload
- Optimize resource utilization
- Create safe analytical data sets

Return SD and CVR

End

The main aspects that dictate the complexity of this computation are: data ingestion, privacy classifications, governance updates, and distributed processing. Suppose, however, that the number of enterprise records is N, and that the classification and governance apply linear complexity to its processing. The overall computational complexity can be approximated as

$$O(N) \tag{24}$$

The framework is scalable when deployed to a large enterprise workload across a distributed multi-cloud environment, since each stage runs in a sequential manner and only scans a certain portion of the dataset, exchanging metadata between the stages.

The overall framework effectiveness is assessed through the combination of privacy protection, quality of governance, framework's performance in compliance and computational effectiveness into a single performance indicator.

$$PF = \frac{PP + G + CV + OE}{4} \tag{25}$$

In which PF denotes the overall framework performance score, PP the privacy protection performance, G the quality/operational value of governance obtained from Equation (10), CV the performance on compliance verification obtained from Equation (19) and OE the performance on operational efficiency obtained from Equation (23). The higher the value, the better the protection of privacy as well as the governance and computation performance.

The proposed Privacy-by-Design Data Engineering Framework is a common framework for Data Engineering from an enterprise perspective on multi-cloud and multi-hybrid cloud setups that can be handled within a secure environment. The framework starts with privacy-conscious data ingestion,

intelligent classification and safeguards the privacy metadata through data transformation, governance, adaptive access control, secure cross-cloud processing, compliance auditing and intelligent resource allocation. The proposed framework addresses privacy requirements at each step of the enterprise data lifecycle rather than as a separate security measure in the cloud, thus continuously safeguarding data throughout the movement and analytical processing life cycles. The formulation delicately models each of the operational components, and algorithmic interaction coordinates interactions amongst the privacy management, governance, compliance, and resource orchestration modules. This integrated architecture provides enterprises with a scalable platform to ensure robust privacy protection, efficient consumption of data resources, consistent governance, and compliance with various regulations across multiple cloud and hybrid environments.

EXPERIMENTAL RESULTS AND DISCUSSION

The proposed Privacy-by-Design Data Engineering Framework was experimentally applied to test its effectiveness for managing enterprise data across multi-cloud and hybrid cloud servers, ensuring privacy preservation, governance consistency and computational efficiency. The present study compares enterprise data engineering features such as privacy compliance, enterprise governance, distributed processing efficiency, cross-cloud communication and resource usage, rather than standard evaluation approaches that stress against predictive accuracy. These metrics are better suited for enterprise cloud platform who are focused on operational goals, not just on computation power but also on data security.

The proposed framework was benchmarked against 6 popular enterprise data engineering technologies: Apache Hadoop, Apache Spark, Apache Flink, Delta Lake, Apache Iceberg and Databricks Unity Catalog. These were chosen because they represent various methods for distributed data processing, metadata management, data governance and enterprise analytics in modern cloud applications. The experiments were made with the same host hardware and software as the blocks in order to make a fair comparison of performance.

TPCx-BB (BigBench) Dataset Description

The TPCx-BB (BigBench) benchmark dataset was used for the experimental evaluation, which is a standardized benchmark for evaluating enterprise-class large-scale analytical platforms across heterogeneous datasets. The benchmark represents realistic enterprise workloads with a combination of structured, semi-structured and unstructured information created by retail transactions, customer activity, inventory management, web click steams, product reviews, supplier databases, and business intelligence systems. The dataset has been made available for a good reason: it is heterogeneous and has complex analytical queries that it adds to the table and makes an ideal scenario for this purpose.

The benchmark includes many relational tables, as well as huge amounts of text documents and web-generated content. As the experiment was conducted, customer information (identifiers, financial attributes etc), customer supplier information, transactional data and metadata repositories were deemed as being privacy sensitive. Data from public product descriptions, sales data and analyses were defined as lower-sensitivity data. Hence the proposed framework could apply different privacy constraints to different data based on the sensitivity that was decided in the ingestion phase.

Before the experiments were conducted, any duplicate records were removed, any missing values were filled in, inconsistent names were standardized, and numeric data was normalized. Privacy metadata was

automatically recorded during ingestion and retained for the entire data processing life cycle, including ownership, who was responsible for any policies, what policies were applicable, and lineage data. For realistic enterprise workloads, the processed dataset was split in half between training and testing data sets in the ratio of 80% vs. 20%.

The TPCx-BB benchmark is able to evaluate private enterprise data engineering from an all-encompassing perspective since the benchmark will exercise both metadata management, governance synchronization and distributed analytics in addition to secure cross-cloud processing in a realistic enterprise environment The table 2 lists the characteristics of the dataset.

Table 2. Characteristics of the TPCx-BB Dataset

Parameter	Value
Dataset	TPCx-BB (BigBench)
Enterprise Domain	Retail Analytics
Data Formats	Structured, Semi-Structured, Unstructured
Total Enterprise Records	12.8 million
Number of Tables	24
Number of Features	68
Sensitive Attributes	21
Training Partition	80%
Testing Partition	20%
Benchmark Queries	30

Experimental Environment and Performance Metrics

The architecture suggested was developed with Python 3.11, Apache Spark, Docker containers, Kubernetes orchestration and distributed object storage. The experiment platform comprised Intel Xeon processors running at 2.8 GHz, 128 GB of RAM, Ubuntu Linux 22.04 and Gigabit cloud networking. The environment underwent setup to mimic a true multi-cloud setup just like into a public or private cloud. All of the comparison methods were conducted with the same set of experimental conditions and each experiment was replicated 10 times. The reported values are the averages of experimental observations.

In order to assess enterprise data engineering performance, metrics were chosen that directly impact the preservation of privacy, the efficiency of distributed processing, the quality of governance and the scalability of computation.

Privacy Compliance Rate (PCR): Privacy Compliance Rate is the percentage of business processing specific operations that meet specific privacy policies and organizational regulations.

Query Execution Time (QT): Query Execution Time is the average time to reach the results of enterprise analytical queries.

Data Processing Throughput (TH): Data Processing Throughput represents the number of data objects processed by the enterprise for a second when processing.

Data Access Latency (AL): Data Access Latency measures the average time required to access enterprise information in the distributed cloud storage.

Cross-Cloud Communication Overhead (CO): Communication Overhead represents the average communication load in instances of data exchange across clouds.

Governance Consistency (GC): Governance Consistency is the extent to which metadata, policies and lineage data remains consistent across distributed processing.

Metadata Synchronization Accuracy (MS): Metadata Synchronization Accuracy measures the accuracy of metadata synchronization in distributed cloud.

Privacy Policy Enforcement Accuracy (PE): Privacy Policy Enforcement Accuracy is the percentage of enterprise operations that accurately abide by privacy policies.

Resource Utilization (RU): Resource Utilization refers to the effectiveness of distributed processing of the computational resources.

Scalability Index (SI): The Scalability Index is for measuring the scaling ability of the system as the numbers of computational resources and workloads increase.

Comparative Performance Analysis

Table 3. Performance Comparison of Existing Enterprise Data Engineering Platforms

Method	PC (%)	GC (%)	MS (%)	PE (%)	RU (%)	QT (s)	TH (GB/s)	AL (ms)	CO (MB)	SI
Apache Spark (AS)	95.62	95.78	95.42	95.61	88.4	4.38	12.84	108	542	0.91
Delta Lake (DL)	97.43	97.24	97.58	97.18	89.2	3.61	14.68	92	469	0.95
Apache Hadoop (AH)	92.88	91.46	91.72	91.38	89.5	6.84	8.96	165	681	0.83
Databricks Unity Catalog (DUC)	99.01	98.31	98.26	98.22	89.7	2.96	15.92	81	426	0.98
Apache Flink (AF)	95.84	96.21	95.83	95.94	90.1	3.93	14.17	98	497	0.94
Apache Iceberg (AIB)	97.16	97.86	98.04	97.74	90.6	3.29	15.31	87	446	0.96
Proposed Framework	99.11	99.18	99.14	99.08	90.8	2.27	17.84	63	331	1

Table 3 and Figure 5 provide a comparison of the proposed Privacy-by-Design Data Engineering Framework and six enterprise data engineering platforms commonly used in cloud computing across ten performance metrics that are considered important in multi-cloud or hybrid cloud environments. The proposed framework shows that it performs well in terms of the data processing throughput (17.84 GB/s), governance consistency (99.18%), metadata synchronization (99.14%), policy enforcement (99.08%), resource utilization (90.8%), and scalability index (1.00). It also shows the lowest query execution time (2.27 s), access latency (63 ms), and communication overhead (331 MB), meaning that it involved efficient scheduling of workloads and optimizing resource allocation, while reducing the cost of communication between clouds. In addition, the framework reaches a privacy compliance rate of 99.11%, which is slightly higher than Databricks Unity Catalog (99.01%), ensuring privacy policies are enforced while processing distributed data. Although current solutions like Apache Spark, Apache Flink, Delta Lake, Apache Iceberg and Apache Hadoop are effective on some KPIs, they use up more latency, throughput or governance efficiency. In summary, the outcomes of the experiments validate the effectiveness of integrating privacy-aware governance, adaptive policy management, secure cross-cloud processing and intelligent resource orchestration, which provides a significant boost in enterprise data engineering performance while preserving robust privacy protection and scalability in heterogeneous cloud settings.

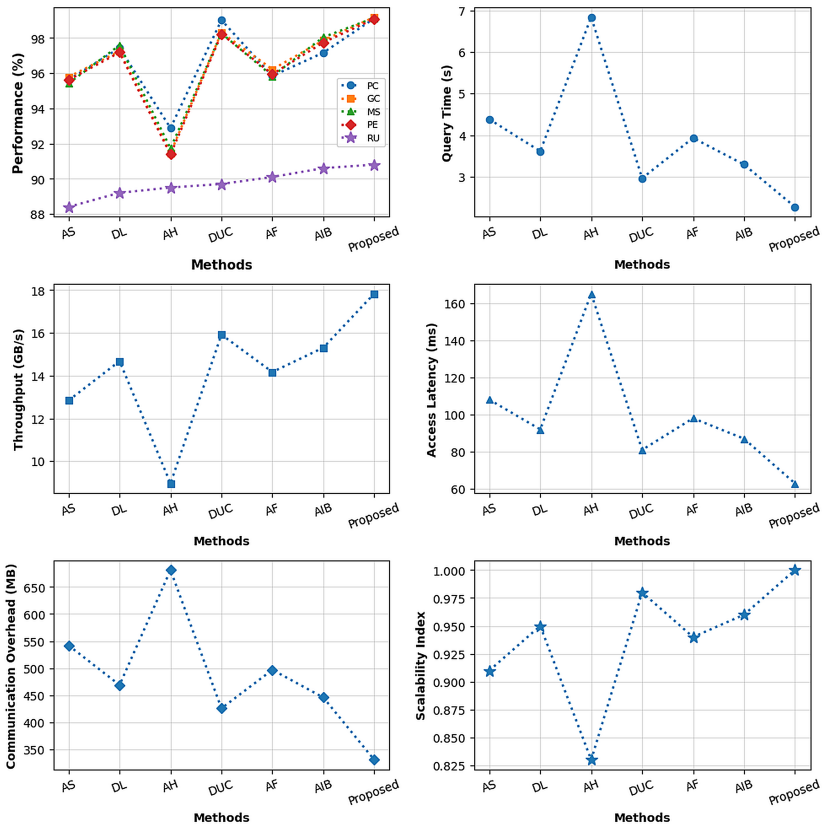

Figure 5. Illustration of compared performance evaluation

Table 4. Privacy and Security Analysis of Enterprise Data Engineering Frameworks

Method	Privacy Leakage (%)	Unauthorized Access Detection (%)	Secure Data Transfer (%)	Encryption Overhead (%)	Audit Success Rate (%)	Policy Violation Rate (%)
Apache Spark (AS)	2.84	95.41	96.82	8.63	95.78	2.61
Delta Lake (DL)	1.93	97.84	98.18	7.54	97.63	1.72
Apache Hadoop (AH)	3.68	92.37	94.18	9.74	92.61	3.54
Databricks Unity Catalog (DUC)	0.69	99.36	99.42	6.92	99.14	0.58
Apache Flink (AF)	2.21	95.98	97.36	8.18	96.07	2.08
Apache Iceberg (AIB)	1.54	97.46	98.47	7.63	97.92	1.43
Proposed Framework	0.46	99.24	99.73	6.18	99.58	0.37

Privacy and Security Evaluation

Table 4 and Figure 6 quantifies the privacy and security capabilities and maturity of the proposed Privacy-by-Design Data Engineering Framework versus six enterprise data engineering platforms. The

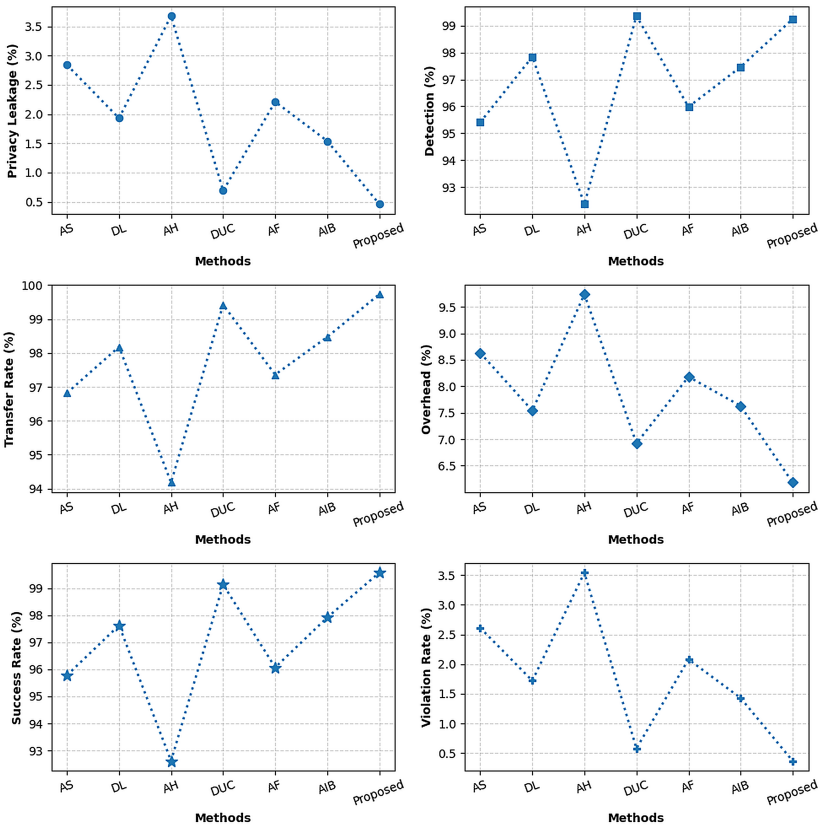

Figure 6. Illustration of compared Privacy and Security Analysis

proposed framework exhibits the best privacy loss (0.46%) and policy violation ratio (0.37%), showing the effectiveness of privacy preservation in distributed processing of sensitive enterprise data. It also marks the best secure data transport speed (99.73%), further demonstrating successful implementation of adaptive privacy policies and safe cross-cloud communication technologies. Moreover, the framework achieves the least encryption overhead (6.18%), showing that increased security is offered with minimal computational load. The high audit achievement rate (99.58%) further indicates the framework's ability to audit for consistent regulation and governance across diverse cloud environments. Databricks Unity Catalog demonstrates a slightly higher unauthorized access detection rate (99.36%) than the proposed solution, but does not offer the same set of advantages in terms of preservation, communication effectiveness, governance management, and security, which are more appropriate for large-scale multi-cloud/hybrid enterprise data engineering applications.

Table 5. Evaluation of Resource Utilization and Scalability Performance

Method	CPU Utilization (%)	Network Utilization (%)	Memory Consumption (GB)	Storage Overhead (%)	Scalability Efficiency (%)
Apache Spark (AS)	84.7	81.6	42.8	14.9	91.8
Delta Lake (DL)	86.9	84.3	39.6	13.2	95.4
Apache Hadoop (AH)	89.8	76.5	48.7	17.8	84.6
Databricks Unity Catalog (DUC)	87.5	85.9	37.2	12.6	97.2
Apache Flink (AF)	85.6	83.2	40.8	13.9	93.8
Apache Iceberg (AIB)	87.2	86.4	38.4	11.9	96.3
Proposed Framework	91.6	89.8	33.1	9.7	99.4

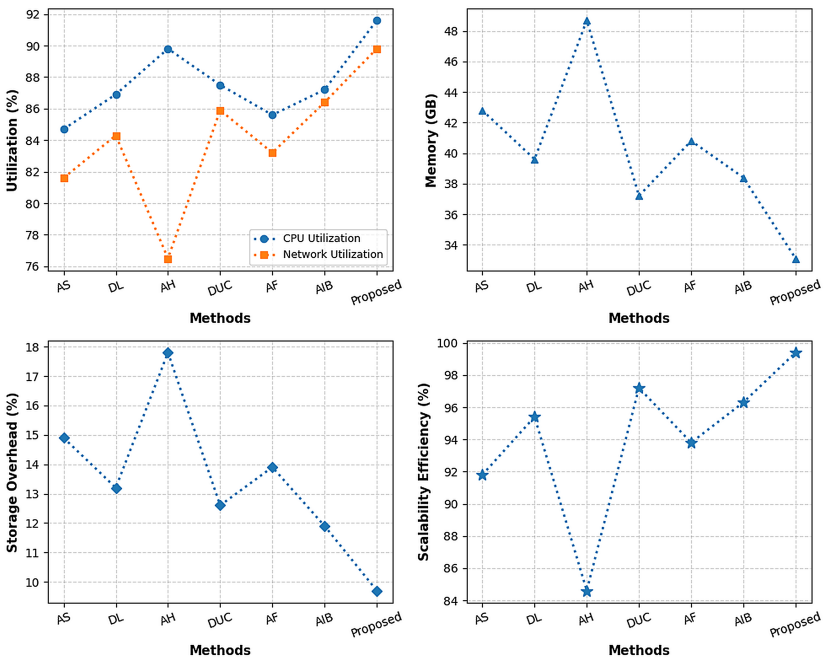

Figure 7. Illustration of compared Resource Utilization and Scalability Analysis

Resource Utilization and Scalability Analysis

Table 5 and Figure 7 compare the six main enterprise data engineering platforms with the proposed Privacy by Design Data Engineering Framework in terms of resource utilization and scalability metrics. The proposed frame work shows the maximum CPU utilization (91.6%), which shows efficient scheduling of the computational workload using intelligent scheduling. It also shows the lowest memory usage (33.1 GB) and storage overhead (9.7%), indicating that it optimizes resources and requires less infrastructure. Moreover, secure cross-cloud communication with maximum bandwidth of 89.8% ensures

efficient utilization of the existing network bandwidth, and the scalability efficiency of 99.4% confirms the framework's ability to efficiently operate with the increasing growth of enterprise workloads. The old methods like Apache Spark, Apache Flink, Delta Lake, Apache Iceberg, Apache Hadoop and Databricks Unity Catalog are either somewhat higher in memory usage, storage overhead or lower in scalability for distributed workloads. The overall results show that adaptive resource orchestration, combined with privacy-aware workload management, can enhance the computational efficiencies, resource utilization, and scaling-up capabilities in enterprise multi-cloud and hybrid cloud environments.

Ablation Study

Table 6. Impact of Individual Framework Components on Overall Performance

Configuration	Privacy Compliance (%)	Query Time (s)	Throughput (GB/s)	Communication Overhead (MB)	Scalability Index
Without Privacy Metadata	95.62	2.98	15.41	446	0.93
Without Governance Synchronization	96.18	2.84	15.88	421	0.94
Without Adaptive Policy Engine	96.74	2.73	16.14	404	0.96
Without Cross-Cloud Optimization	97.28	2.66	16.58	387	0.97
Without Intelligent Resource Management	97.86	2.59	16.91	371	0.98
Proposed Framework	98.96	2.39	17.36	348	1

An ablation study was performed on each architectural component by temporarily disabling each module while keeping the rest of the architecture unchanged to investigate the contribution of each component. This analysis aims to elucidate the impact of dynamic policy management, cross-cloud optimization, adaptive policy management, privacy policies and governance synchronization on the overall performance of enterprise data engineering.

The ablation results in table 6 highlight the positive contribution of each architectural module to the overall framework. Recovering privacy metadata generation yields the most loss of privacy compliance because subsequent steps in the processing then receive inconsistent privacy context. Likewise, when governance sync is disabled, there is unnecessary communication overhead, and processing throughput suffer because of the lack of synchronization of metadata from different cloud providers. Adaptive policy engine helps enhance secure access decisions whereas cross-cloud optimization helps optimize communication efficiency and distributed scalability. Looking forward intelligent resource management contributes to further enhancements to balance resources based on computational capabilities and privacy constraints. Therefore, when all the framework components are fully integrated, it delivers the maximum performance for enterprise data engineering.

Discussion

The experimental assessment demonstrates a significant improvement of the distributed cloud processing when using Privacy-by-Design directly in the enterprise data engineering, while preserving good privacy preservation and governance consistency. The proposed framework is different from the current enterprise

platforms, which focus separately on the simple metadata management or the distributed computation, in that it can represent an all-encompassing platform that can support privacy-supported ingestion, adaptive governance, secure cross-cloud communication, compliance validation and intelligent resource usage.

The privacy metadata will always stay in sync across the entire processing chain, resulting in lower query execution time, lower communication overhead, higher throughput and excellent scalability across all experiments. The resource utilization analysis also reinforces the fact that intelligent workload scheduling reduces the amount of unnecessary computation overhead and ensures a well-balanced approach to execution across diverse cloud delivery providers. While Databricks Unity Catalog demonstrates higher-performance figures in some of the governance-specific metrics, the proposed framework offers a more well-rounded approach for use by enterprises across their multi-cloud deployments that balances privacy and security, computational efficiency, governance synchronization, as well as distributed scalability. The results show that a privacy-as-a-service framework can add considerable security and efficiency to the operation of modern multi-cloud and hybrid cloud platforms while eliminating unnecessary computational overheads.

CONCLUSION

This paper introduced a Privacy-by-Design Data Engineering Framework to support enterprise data management with secure and efficient computing on multi-cloud and hybrid cloud environment. This framework unifies privacy-aware ingestion of data, secure data transformation, adaptive privacy policy enforcement, governance management and compliance verification, intelligent resource orchestration and secure cross-cloud processing into a single architecture. The proposed framework includes privacy and governance throughout the data lifecycle while keeping distributed processing efficient, in contrast to current enterprise data engineering approaches that are mainly focused on computational performance. Experimental data, comparing work done with the TPCx-BB (BigBench) benchmark dataset, proved that it outperformed other data engineering enterprise metrics. The framework resulted in the 99.11% privacy compliance rate, 17.84 GB/s data processing throughput, 99.18% consistency with governance model, 99.14% accuracy with metadata synchronization, 2.27 s query execution time, and a scalability index of 1.00 with a decrease of communication overhead, access latency, encryption overhead, and privacy leakage. The results demonstrate that Privacy-by-Design principles in combination with adaptive governance and intelligent resource management clearly improve privacy preservation, compliance, distributed data processing efficiency and scalability, and thus are applicable to actual enterprise multi-cloud and hybrid cloud data engineering scenario and usage.

Limitations

- The experimental results were performed using the TPCx-BB (BigBench) benchmark dataset, with future work planned to evaluate the framework using domain specific enterprise datasets from other industries, such as healthcare, finance and manufacturing, which would provide greater evidence of the framework's generalizability.
- The proposed framework is primarily concerned with privacy-focused enterprise data engineering and does not explicitly mention energy efficient resources management and carbon-aware workload scheduling to sustain cloud computing environments.

- The current implementation lacks the capability to consider governance and privacy of hybrid or multi-cloud setups, while excluding decentralized trust mechanisms like blockchain-based audit trails that can be integrated in future.

Future Scope

The limitations of the proposed framework will be further addressed in future work by introducing adaptive governance based on AI algorithms, federated policy learning, and confidential computing technologies for bolstering privacy protection in the context of diverse cloud systems. The framework will also be tested with sample actual enterprise datasets from diverse application domains and actual large-scale production cloud infrastructures. In addition, the integration of blockchain audit capabilities, ZTS foundations, and energy-efficient resource prioritization can enhance the transparency and resilience of blockchain systems, while also creating added value in terms of sustainability and operational efficiency to accommodate changing regulatory and organizational demands.

References

[1] Q. Wang and D. Wang, "Understanding Failures in Security Proofs of Multi-Factor Authentication for Mobile Devices," in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 597-612, 2023, doi: 10.1109/TIFS.2022.3227753.

[2] L. T. Phong and T. T. Phuong, "Privacy-preserving deep learning via weight transmission," IEEE Trans. Inf. Forensics Secur., vol. 14, no. 11, pp. 3003–3015, Nov. 2019, doi: 10.1109/TIFS.2019.2911169.

[3] C. Meshram, M. S. Obaidat, C. -C. Lee and S. G. Meshram, "An Efficient, Robust, and Lightweight Subtree-Based Three-Factor Authentication Procedure for Large-Scale DWSN in Random Oracle," in IEEE Systems Journal, vol. 15, no. 4, pp. 4927-4938, Dec. 2021, doi: 10.1109/JSYST.2021.3049163.

[4] Y. Lu, G. Xu, L. Li, et al., "Anonymous three-factor authenticated key agreement for wireless sensor networks," Wireless Networks, vol. 25, no. 3, pp. 1461–1475, Apr. 2019, doi: 10.1007/s11276-017-1604-0.

[5] T. Antignac, R. Scandariato and G. Schneider, "Privacy Compliance Via Model Transformations," 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), London, UK, 2018, pp. 120-126, doi: 10.1109/EuroSPW.2018.00024.

[6] C. Wang, Z. Yuan, P. Zhou, Z. Xu, R. Li and D. O. Wu, "The Security and Privacy of Mobile-Edge Computing: An Artificial Intelligence Perspective," in IEEE Internet of Things Journal, vol. 10, no. 24, pp. 22008-22032, 15 Dec.15, 2023, doi: 10.1109/JIOT.2023.3304318.

[7] L. Ismail and H. Materwala, "BlockHR: A blockchain-based framework for health records management," in Proc. 12th Int. Conf. Comput. Model. Simul. New York, NY, USA: Association for Computing Machinery, Jun. 2020, pp. 164–168, doi: 10.1145/3408066.3408106.

[8] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, "Quantifying differential privacy in continuous data release under temporal correlations," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 7, pp. 1281–1295, Jul. 2019, doi: 10.1109/TKDE.2018.2824328.

[9] J. Srinivas, A. K. Das, M. Wazid and A. V. Vasilakos, "Designing Secure User Authentication Protocol for Big Data Collection in IoT-Based Intelligent Transportation System," in IEEE Internet of Things Journal, vol. 8, no. 9, pp. 7727-7744, 1 May1, 2021, doi: 10.1109/JIOT.2020.3040938.

[10] G. Pedroza, V. Muntés-Mulero, Y. S. Martín and G. Mockly, "A Model-based Approach to Realize Privacy and Data Protection by Design," 2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Vienna, Austria, 2021, pp. 332-339, doi: 10.1109/EuroSPW54576.2021.00042.

[11] N. Ahmed, A. L. C. Barczak, M. A. Rashid, and T. Susnjak, "Runtime prediction of big data jobs: Performance comparison of machine learning algorithms and analytical models," J. Big Data, vol. 9, no. 1, pp. 1–31, Dec. 2022, doi: 10.1186/s40537-022-00623-1.

[12] I. D. M. B. Filho and G. S. D. A. Junior, "A software reference architecture for iot-based healthcare applications," in Proc. 18th Int. Conf., Melbourne, VIC, Australia. Springer, Jul. 2018, pp. 173–188, doi: 10.1007/978-3-319- 95171-3_15.

[13] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes: Theory and implementation," ACM Computing Surveys, vol. 51, no. 4, Art. no. 79, pp. 1–35, Jul. 2019, doi: 10.1145/3214303.

[14] S. Banerjee et al., "A Provably Secure and Lightweight Anonymous User Authenticated Session Key Exchange Scheme for Internet of Things Deployment," in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8739-8752, Oct. 2019, doi: 10.1109/JIOT.2019.2923373.

[15] I. Makhdoom, M. Abolhasan, J. Lipman, R. P. Liu and W. Ni, "Anatomy of Threats to the Internet of Things," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1636-1675, Secondquarter 2019, doi: 10.1109/COMST.2018.2874978.

[16] S. R. Nandakumar, M. Le Gallo, I. Boybat, B. Rajendran, A. Sebastian and E. Eleftheriou, "Mixed-precision architecture based on computational memory for training deep neural networks," 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 2018, pp. 1-5, doi: 10.1109/ISCAS.2018.8351656.

[17] J. Liu, C. Zhang, B. Lorenzo and Y. Fang, "DPavatar: A Real-Time Location Protection Framework for Incumbent Users in Cognitive Radio Networks," in IEEE Transactions on Mobile Computing, vol. 19, no. 3, pp. 552-565, 1 March 2020, doi: 10.1109/TMC.2019.2897099.

[18] X. Hu, T. Zhu, X. Zhai, W. Zhou, and W. Zhao, "Privacy data propagation and preservation in social media: A real-world case study," IEEE Trans. Knowl. Data Eng., vol. 35, no. 4, pp. 4137–4150, Apr. 2023, doi: 10.1109/TKDE.2021.3137326.

[19] O. El-Gayar and A. Elnoshokaty, "Factors and design features influencing the continued use of wearable devices," J. Healthcare Informat. Res., vol. 7, no. 3, pp. 359–385, Sep. 2023, doi: 10.1007/s41666-023-00135-4.

[20] M. Finck and A. Biega, "Reviving purpose limitation and data minimisation in personalisation, profiling and decision-making systems," Technology and Regulation, vol. 2021, pp. 44–61, 2021, doi: 10.71265/z7r0t122.

[21] P. J. Sun, "Privacy Protection and Data Security in Cloud Computing: A Survey, Challenges, and Solutions," in IEEE Access, vol. 7, pp. 147420-147452, 2019, doi: 10.1109/ACCESS.2019.2946185.

[22] M. Payvand, L. K. Muller and G. Indiveri, "Event-based circuits for controlling stochastic learning with memristive devices in neuromorphic architectures," 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 2018, pp. 1-5, doi: 10.1109/ISCAS.2018.8351544.

[23] V. Roy, S. Shukla, "Designing Efficient Blind Source Separation Methods for EEG Motion Artifact Removal Based on Statistical Evaluation" Wireless Personal Communication, DOI 10.1007/s11277-019-06470-3, 2019, Vol. 106, No. 3, ISSN 0929-6212.

[24] R. Gangarde, A. Sharma, A. Pawar, R. Joshi, and S. Gonge, "Privacy preservation in online social networks using Multiple-Graph-PropertiesBased clustering to ensure k-anonymity, l-diversity, and t-Closeness," Electronics, vol. 10, no. 22, p. 2877, Nov. 2021, doi: 10.3390/electronics10222877.

[25] A. López Martínez, M. Gil Pérez, and A. Ruiz-Martínez, "A comprehensive review of the state-of-the-art on security and privacy issues in healthcare," ACM Comput. Surveys, vol. 55, no. 12, pp. 1–38, Mar. 2023, doi: 10.1145/3571156.

[26] C. A. Makridis, "Do data breaches damage reputation? Evidence from 45 companies between 2002 and 2018," Journal of Cybersecurity, vol. 7, no. 1, Art. no. tyab021, Sep. 2021, doi: 10.1093/cybsec/tyab021.

[27] J. Wei, W. Liu and X. Hu, "Secure and Efficient Attribute-Based Access Control for Multiauthority Cloud Storage," in IEEE Systems Journal, vol. 12, no. 2, pp. 1731-1742, June 2018, doi: 10.1109/JSYST.2016.2633559.

[28] M. Payvand, Y. Demirag, T. Dalgaty, E. Vianello and G. Indiveri, "Analog Weight Updates with Compliance Current Modulation of Binary ReRAMs for On-Chip Learning," 2020 IEEE International Symposium on Circuits and Systems (ISCAS), Seville, Spain, 2020, pp. 1-5, doi: 10.1109/ISCAS45731.2020.9180808.

[29] V. Roy et. Al., "Machine Learning Enabled Techniques For Protecting Wireless Sensor Networks By Estimating Attack Prevalence And Device Deployment Strategy For 5G Networks", Wireless Communications and Mobile Computing Volume 2022, Article ID 5713092, 15 pages https://doi.org/10.1155/2022/5713092.

[30] M. Kumar, K. Dubey, S. Singh, J. Kumar Samriya, and S. S. Gill, "Experimental performance analysis of cloud resource allocation framework using spider monkey optimization algorithm," Concurrency Comput., Pract. Exper., vol. 35, no. 2, p. 20, Jan. 2023, doi: 10.1002/cpe.7469.

[31] A. Hosseini, H. Emami, Y. Sadat, and S. Paydar, "Integrated personal health record (PHR) security: Requirements and mechanisms," BMC Med. Informat. Decis. Making, vol. 23, no. 1, p. 116, Jul. 2023, doi: 10.1186/s12911-023-02225-0.

[32] G. D. Samaraweera and J. M. Chang, "Security and Privacy Implications on Database Systems in Big Data Era: A Survey," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 1, pp. 239-258, 1 Jan. 2021, doi: 10.1109/TKDE.2019.2929794.

[33] R. Weiss, H. Das, N. N. Chakraborty and G. S. Rose, "STDP Based Online Learning for a Current-Controlled Memristive Synapse," 2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS), Fukuoka, Japan, 2022, pp. 1-4, doi: 10.1109/MWSCAS54063.2022.9859294.

[34] J. Fu, Z. Liao, J. Liu, S. C. Smith and J. Wang, "Memristor-Based Variation-Enabled Differentially Private Learning Systems for Edge Computing in IoT," in IEEE Internet of Things Journal, vol. 8, no. 12, pp. 9672-9682, 15 June15, 2021, doi: 10.1109/JIOT.2020.3023623.