Global-scale GDPR Compliant Data Sharing System

Yeonghun Nam, Eunok Shin, Suyeong Lee, Seungho Jung, Yohan Bae, and Junghyun Kim
Samsung Research
Samsung Electronics
Seoul, Republic of Korea
{yeonghun.nam, eunok.shin. sy710.lee, shonest.jung, yhan.bae, jh24.kim}@samsung.com

***Abstract*—Personal data protection associated with regulatory compliance is often siloed as legal documentation that is not suitable for GDPR compliant system implementation. In a real global-scale industrial environment, applying novel personal data protection technologies directly to industrial sites has the potential risk to cause unintended disclosure of personal information, which results in a huge monetary penalty. In this paper, the Global-scale GDPR Compliant Data Sharing (GDS) system is introduced which enables efficient data retrieval and data sharing process while protecting personal information.**

Keywords— Data Privacy*; *Data Sharing*; *Big Data*; *GDPR

I. INTRODUCTION

Data protection regulatory technologies have long existed, and evolved along with the society and technology. The General Data Protection Regulation (GDPR) [1] defines data protection regulatory, data subjects' rights, and legal obligations so as to ensure the protection of EU citizens' personal information. Those regulatory innovations do have an impact on the technological products that must abide to them, and on the engineering process followed for system development. For instance, products must implement any functionality needed to support GDPR compliance for data subject to enforce their rights.

Personal data protection associated with regulatory compliance is often siloed as legal documentation that is not suitable for real system implementation. There are several novel data privacy technologies, however, they are isolated from the data protection regulatory such as GDPR. Moreover, in a real global-scale industrial environment, applying new data protection technologies directly to industrial sites has the potential risk to cause unintended disclosure of personal information. The GDPR mandates penalties of up to € 20 million or 4 percent of the global annual turnover for non-compliance. Therefore, in a real industrial environment, for all data sharing processes, not just EU citizens' personal information sharing process, it is essential to implement a system architecture that is fully GDPR compliant while safely utilizing personal information.

In this paper, the Global-scale GDPR Compliant Data Sharing (GDS) system is proposed which enables efficient data retrieval and data sharing process while protecting personal information. The system architecture is designed that enables efficient data retrieval and data sharing process while protecting personal information. The GDS system also includes the methodology to manage regulatory as metadata: legal obligation, data usage consent for data subject, type of personal information in each dataset, data processing agreement between organizations. Finally, global-scale GDPR compliant data sharing process is proposed.

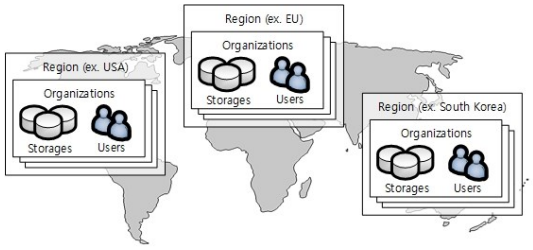

Figure 1. Data stored in different regions

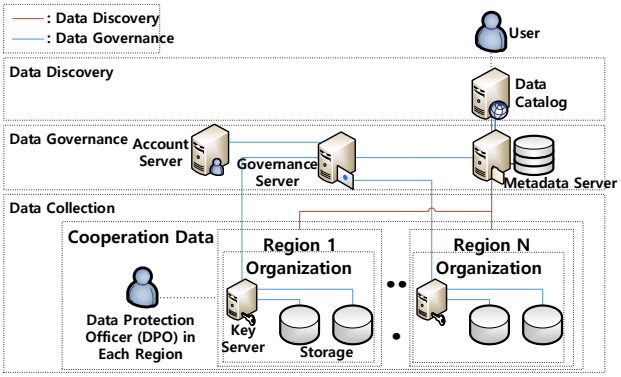

Figure 2. Overall Deployment Architecture

II. GDS ARCHITECTURE

In this Section, The GDS architecture is proposed, and the metadata to be managed by each entity is designed. The system overview and data sharing requirements are described in Section II-A, and the functionalities of proposed system layers are described in Section II-B, II-C, and II-D.

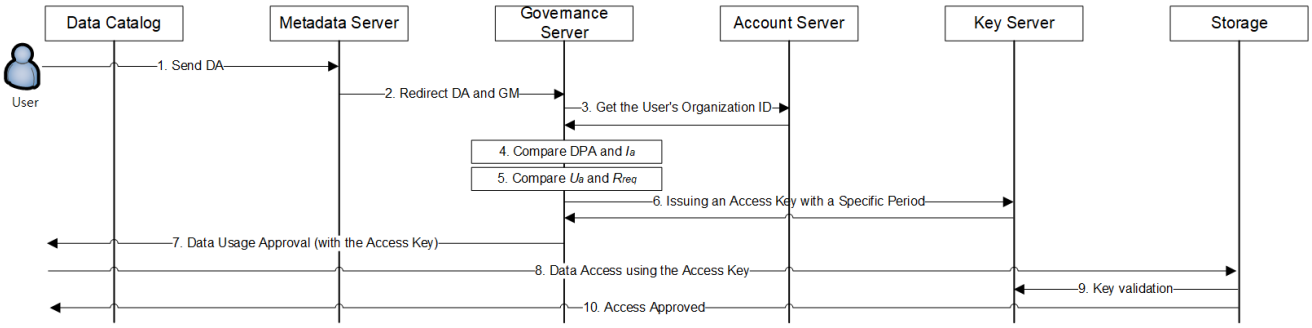

Figure 3. Data Usage Process.

A. System Overview

As presented in Fig. 1., an organization can provide services to its region, and stores data in a data center. Each organization analyzes stored data to predict product demand, detect anomalies, or improve the user experience. At this time, the organization must inform the user of the policy regarding the storage and use of data, and data must be processed only within the range agreed by each user. When global-scale data analysis is being conducted, data distributed across multiple regions needs be delivered to a specific region. Based on GDPR, each organization cannot move personal information to a specific region without the appropriate data sharing process. To determine the validity of data sharing between organizations, the Data Processing Agreements (DPAs) [3] must be concluded, which includes the data processing instructions, duty of confidentiality, technical and organizational measures for data security, and so on.

Therefore, user's data usage consent and DPA must be considered for data sharing between organizations in different regions. The overall deployment architecture of GDS is presented in Fig. 2, where the system can be divided into data collection, data governance, and data discovery layers.

B. Data Collection Layer

The *Data Collection Layer* is a layer that stores user data. There are two kinds of entities in the *Data Collection Layer*: storages server and a key server. A storage (including storage management server) server is the entity in which the data is stored. Personal information is stored and managed in each organization's storages where only data consented by data subjects can be collected in the storage. Each organization that collects data plays a role of the data controller, and before new data is collected, the Data Protect Officer (DPO) [2] appointed by region or organization creates a Governance Metadata (GM) for the data to be collected. Then the GM it is delivered to the Metadata Server in the Data Governance Layer (see in Section 2-3). For the dataset ID d_a, GM format can be defined as $\{d_a, I_a, U_a, o_a\}$ where I_a is the personal information matrix, U_a is the consent matrix for data subjects, o_a is the organization ID. The key server issues an access key for the user to access the storage when the data sharing procedure is completed.

C. Data Governance Layer

In the *Data Governance Layer*, metadata collected from the organizations are managed, and the data sharing requests are validated so that only the authorized users can access the data. To determine the validity of data sharing between organizations, the Data Governance Layer manages the DPAs concluded between organizations (The detailed data sharing process is continued in Section 3). There are three kinds of entities in the *Data Governance Layer*: Metadata Server, Account Server, and Governance Server. Metadata Server stores dataset metadata delivered from each organization. Each metadata includes description as well as governance information. The account server manages the user ID, organization ID, and region information in the GDS system. The organization IDs of the data controller and processor stored in the account server are referred to the data sharing process.

The Governance Server is an entity that processes data sharing requests and allows data access with the consent of the data subject. Data must be collected explicitly and legally, and if data processing is incompatible with the original purpose, it should not be processed without consent. Therefore, for a data sharing request, the Governance Server determines if the data subject agrees with the data processing. In addition, it determines whether data sharing is possible based on the DPA concluded between the Data Controller and the Data Processing organizations.

D. Data Discovery Layer

There is a Data Catalog in the *Data Discovery Layer*, which is a metadata management tool designed to help organizations find data based on metadata located in the Metadata Server while real data is stored in each organization storage. It helps data sources more discoverable and manageable for users and helps organizations make more informed decisions about how to use their data. After discovering metadata, each user can start the data sharing process, described in the Section III.

III. GDPR Compliant Data Sharing Process

In this Section, the data sharing process is proposed in the GDS system. In this paper, the data sharing process is divided into three phases: i) Data Sharing Request, ii) Data Sharing

Approval, and iii) Data Use. The data sharing diagram in the GDS system is presented in Fig. 3.

A. Data Sharing Request

In the Data Sharing Request phase, a user retrieves data through the Data Catalog and sends a Data-use Application (DA) request to the Metadata Server. The DA format for the ath dataset is $\{u_{\text{req}}, d_{\text{a}}, R_{\text{req}}, [t_{\text{from}}, t_{\text{to}}]\}$ where u_{req} is the user ID, d_{a} is the dataset ID, R_{req} is the data-use purposes matrix, and $[t_{\text{from}}, t_{\text{to}}]$ is the requested the data-usage period.

B. Data Sharing Approval

In this phase, data usability is determined based on the i) DA from the user, ii) the GM stored in the Metadata Server, and ii) the DPA stored in the Governance Server. When the user's DA arrives at the Metadata Server, the Metadata Server redirects the GM $\{d_a, I_a, U_a, o_a\}$ and DA information $\{u_{\text{req}}, d_{\text{a}}, R_{\text{req}}, [t_{\text{from}}, t_{\text{to}}]\}$ to the Governance Server. In the Governance Server, the DPA validation is conducted at first. The Governance Server retrieves the organization ID o_{req} of the user u_{req} from the Account Server, and loads DPA metadata between o_{req} and $o_{a.}$ Finally, it determines if the personal information in the d_a is outside the scope of the DPA. Next, the it validates if the user's data usage plan is within the data usage scope that the data subject agreed to, by comparing U_a and $R_{\text{req.}}$

C. Dataset Use

When the validation is finished, the Governance Server issues an access token through the organization's key server (where d_a is located) which is valid for the period $[t_{from}, t_{to}]$.

IV. Related Works

There are several technologies that consider GDPR-compliant data usage systems. In [4], necessary entities were defined based on the GDPR document, and the message flow and types among entities were defined. In [5], a distributed personal information tracking system is proposed based on block chain technology. However, these technologies do not include practical requirements and system flows that can be used directly in the real industrial world.

V. Conclusion

The GDPR is becoming the standard that dictates privacy and security for personal information. To meet these requirements, in Apr. 2019, the Data Catalog service is released in the Samsung Research (SR) that enables researchers at SR, Samsung AI Center, and Oversea Research Centers to explore datasets and use dataset based on the GDPR compliant data sharing process.

Future research needs to be extended to apply the novel de-identification and anonymization technologies to the GDS system for data processors in organizations that do not have DPAs to access personal information. The authors plan to continue research on these issues.

References

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1, 2016.

[2] Everything you need to know about the GPDR Data Protection Officer (DPO), GDPR.EU., https://gdpr.eu/data-protection-officer/

[3] What is a GDPR data processing agreement?, GDPR.EU., https://gdpr.eu/what-is-data-processing-agreement/

[4] H. J. Pandit, D. O'Sullivan, and D. Lewins, "GDPR Data Interoperability Model," *23rd EURAS Annual Standardisation Conference*, Jun. 2018.

[5] M. M. H. Onik, C-S. Kim, N-Y. Lee, and J. Yang, "Privacy-aware blockchain for personal data sharing and tracking," Open Computer Scienct, Apr. 2019.