EAI Endorsed Transactions on Security and Safety Research Article EAI.EU

GIP3: Make Privacy Preserving be Easier on Cloud

Shiying Pan[1,*], Can Yang[1], Runmin Li[1]

[1]South China University of Technology

Abstract

The lack of data privacy preserving tools in the cloud is an urgent issue to solve today. To meet the needs of data sharing and data publishing, this paper proposed a cloud-oriented privacy preserving framework, in which we designed and implemented a SaaS data privacy preserving platform, called GIP3 (General Web Data Interface - Privacy Preserving Protocol). This platform supports a variety of mainstream privacy preserving approaches, and users can use them to implement data privacy preserving and evaluate the utility of data with different information loss metrics. This platform integrates with an general web data interface to perform a cloud-oriented multi-tenant data management, which provides a complete data service chain from system construction, collection, management and maintenance, privacy preserving, to publishing. In a word, it makes data privacy preserving be easier on the cloud.

Keywords: Anonymity, Cloud Computing, Data Privacy Preserving SaaS, Web Service.

Received on 28 November 2020, accepted on 10 January 2021, published on 12 January 2021

doi: 10.4108/eai.25-6-2021.170249

1. Introduction

At present, massive of heterogeneous data are migrating from local to cloud, and cloud data sharing can provide powerful support for scientific research and decision. However, there may be a lot of sensitive information in the massive data, and the data publishing without privacy preserving would bring a crisis of privacy disclosure[5]. In order to preserve the sensitive information in the data and ensure the usability of the data, researchers proposed a data publishing technology based on information limitation. This technique focuses on trans- forming raw data into privacy-preserving versions that protect the data owner and his/her sensitive information from disclosure while ensuring the utility of data [1], in which the procedure of transforming data is called data anonymization. In recent years, many researchers have proposed data privacy preserving approaches in different scenarios. However, there is a lack of general web systems for data privacy preserving service for users. In fact, one of the major barriers to developing sustainable and efficient data systems is the lack of reliable and convenient privacy tools [4]. Currently, local client tools cannot meet the needs of users to manage their data in the cloud, but SaaS creates the possibility for it. SaaS is an emerging software application model which is widely used [15], in which tenants can ordering cloud services provided by SaaS providers.

Based on above reasons, we developed the privacy preserving cloud platform based on SaaS, which enables users to use required privacy preserving services on the Web. This platform is based on the SaaS data management system platform, called GWDI (General Web Data Interface) [22], which meets the needs of users to customize data management systems and publish or manage data. In this paper, we focus on the design and implementation of the platform, which further provides privacy preserving service based on GWDI, and is called GIP3 mean ing General Web Data Interface - Privacy Preserving Protocol. GIP3 enables users under different tenants to interactively execute the data anonymization, then publish and manage data in the cloud. It supports a variety of privacypreserving approaches for relational and transactional data, and it also provides different information loss metrics so that users can evaluate data utility and make decisions. In addition, GIP3 provides a unified Web interface in the cloud for users.

* Corresponding author. Email: 396852378@qq.com

The remainder of this paper is organized as follows. Section 2 introduces our related work. Section 3 describes the design and implementation of GIP3 in de- tail. Section 4 evaluates the feasibility and effectiveness of GIP3 by experiments. Finally, Section 5 summarizes the work in this paper and proposes the future work.

2. Related Work

This section investigates related researches and compares them with GIP3.

Xiao X et al. proposed an interactive data anonymization tool called CAT[20], which is a client program implemented in C++. It supports k-anonymity and l-diversity. It also provides a risk assessment for each record, and then provides a method for users to manually suppress the records. Dai C et al. proposed a tool called TIAMAT[3] that supports k-anonymity. It provides discriminability metric(DM) and normalized certainly penalty(NCP) for user to evaluate information loss and data utility. Moreover, UTD-AT(UTD-Anonymization Tool)[7] is a loosely coupled command-line tool implemented in JAVA. It supports k- anonymity, l-diversity, t-closeness, and anatomy. The user configures the privacy model and parameters by an XML file in a specific format. However, it does not have a visual UI, and it further requires the support of SQLite database, which has some shortcomings in scalability. Poulis G et al. proposed a client tool called SECRETA[12] to evaluate different anonymization approaches. It provides four different algorithms such as clustering algorithm and Mondrain algorithm to achieve k-anonymity for relational data, and five different algorithms to achieve km-anonymity for transactional data. It provides NCP for measuring information loss. However, users must be on a Linux system to install and use it. Prasser F et al. implemented a client tool called ARX[14]. It is a relatively mature tool, which supports a variety of privacy preserving approaches such as k-anonymity, l-diversity, t-closeness, etc. It also integrates a variety of information loss metrics such as DM, NCP, etc., in which it also provides risk assessment after anonymization. However, while they plan the anonymization of transactional data in future work, it is still not implemented. μ-argus[6] is a tool that provides a traditional method of anonymization. Users manually set the disclosure risk threshold, and then generalize and suppress data, until the disclosure risk is reduced to the threshold range.

On the other hand, the rise of SaaS also brings the researches of privacy pre- serving in SaaS field. Current researches focus on multi-tenant data isolation, identity-based access control, and privacy detection of user behavior[11][8]. But in this paper, we propose anonymization for user data. According to the investigation, there is still no design or system platform similar to GIP3. Most of the existing privacy preserving tools are client/server architectures, which mainly support local data processing and have some shortcomings in scalability. How- ever, we provide data privacy preserving in the SaaS cloud and integrate the data management system to enable users to access, process and publish data on the Web. In addition, the existing tools support few anonymization approaches, some of them do not provide information loss metrics. So we further enrich the privacy preserving approaches and information loss metrics.

3. Design and Implementation

Table 1 details the features of GIP3 and related open source tools (CAT[20], UTD-AT[7], ARX[14]).

Table 1. Features of different privacy preserving tools

	GIP3	CAT	UTD-AT	ARX
Relational data	X	X	X	X
Transactional data	X			
Relational-Transactional data	X			
Information loss met-rics	X			X
Visualization UI	X	X		X
Web environment	X			
Data management and publishing on SaaS	X			

Table 1 shows that GIP3 not only serves relational data, but also trans- actional data and relational-transactional data. Moreover, it supports different information loss metrics to evaluate data utility. In addition, it has a visual UIfor users on the web, which avoids the lack of cross-platform scalability and in- convenience caused by downloading and installing. More importantly, it inherits GWDI's multi-tenant mode based on SaaS and provides a complete cloud data management system service, in which it integrates the data storage, publishing, maintenance, statistics, and privacy preserving services required by users on the platform. Fig. 1 shows the architecture of the cloud privacy preserving

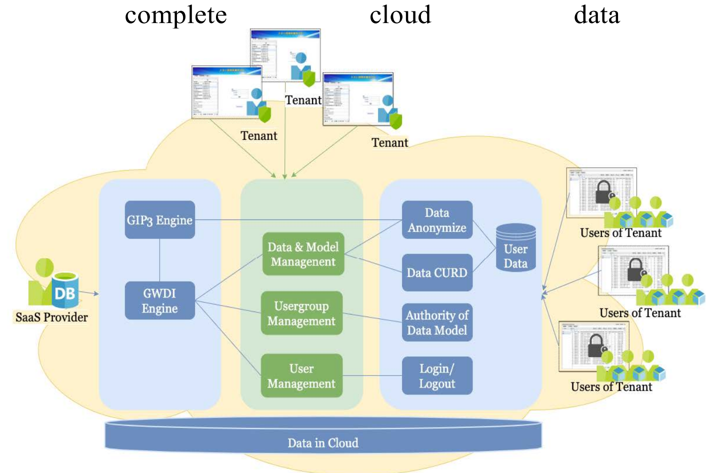


Fig. 1. Overall Architecture of SaaS privacy preserving platform (GIP3)

platform based on GWDI, in which the tenant builds the data management system in GWDI on demand, and the user logs into the tenant's data management system and then manage data, anonymize data and publish data etc.

3.1 Architecture Design

The GIP3 privacy preserving module adopts a loosely coupled design, which is divided into user presentation layer and system processing layer, as shown in Fig. 2.

The user presentation layer provides Internet interface by the web server, and the system processing layer anonymize the users' data by the anonymize server. The user presentation layer is divided into the following three modules.

Data interface module. This module integrates GWDI's data management system so that users can view the data before and after anonymization, and manage the data.

Parameter configuration module. This module is responsible for the selection of privacy model and information loss metric, as well as the configuration of model's parameters, the definition of related data fields (for example, sensitive attribute) etc.

Data processing module. This module is the interface for users to load raw data and export or download anonymous data.

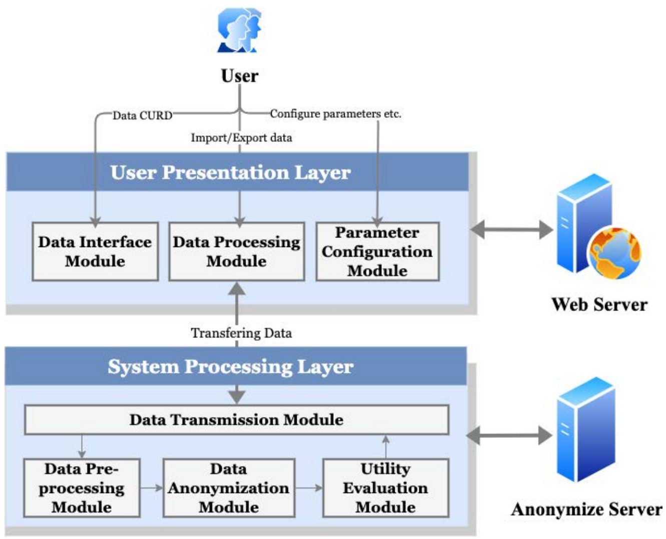


Fig. 2. Software architecture of GIP3 privacy preserving module

The user makes requests in the user presentation layer, and then the anonymization is processing in the system processing layer. The system processing layer is divided into the following four modules.

Data transmission module. This module is responsible for receiving data from the user presentation layer and transferring anonymous data to the user presentation layer.

Data pre-processing module. This module is responsible for reading and pre- processing raw data, in which it currently supports CSV and XLS formats. It deals primarily with missing or error records in raw data.

Data anonymization module. This module is responsible for anonymizing the pre-processed data according to the configuration.

Utility evaluation module. This module is responsible for evaluating the utility of the anonymous data by the configured information loss metric.

According to the architecture design, the procedure of a user implementing data privacy preserving on GIP3 is shown in Fig. 3.

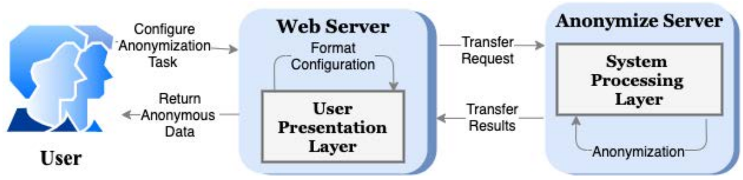


Fig. 3. User implements privacy preserving procedure on GIP3

3.2 Module Implementation

User Presentation Layer The main tasks of the user presentation layer are data import, parameter configuration and data export, in which its core UML is shown in Fig. 4. The user presentation layer mainly uses the AnonymizerMan- ager class to implement various methods, including importing/exporting data, obtaining configuration, etc., in which the AnonymizerModel class represents supported privacy preserving models.

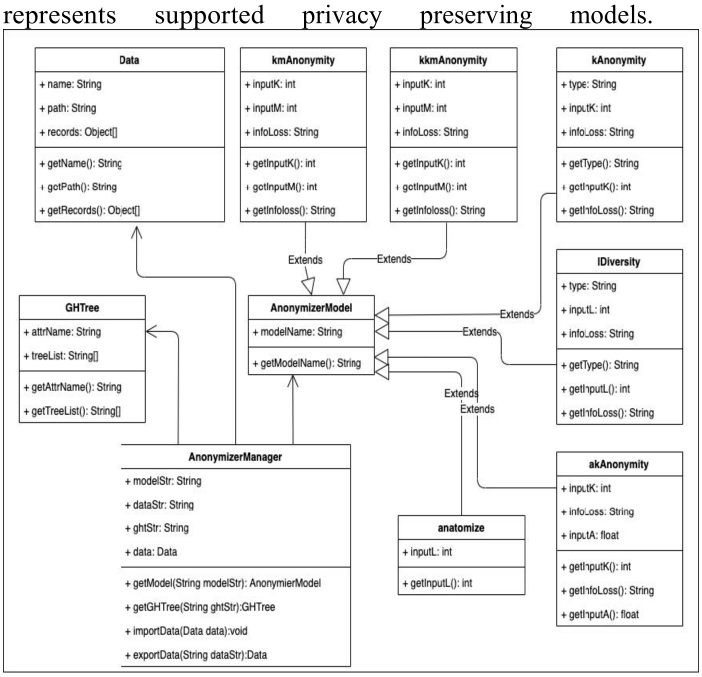


Fig. 4. UML class diagram of User presentation layer core component

Fig. 5 represents a UML sequence diagram of an anonymization request from a user on GIP3.
System Processing Layer The privacy preserving models and information loss metrics provided in the system processing layer are shown in Table 2.

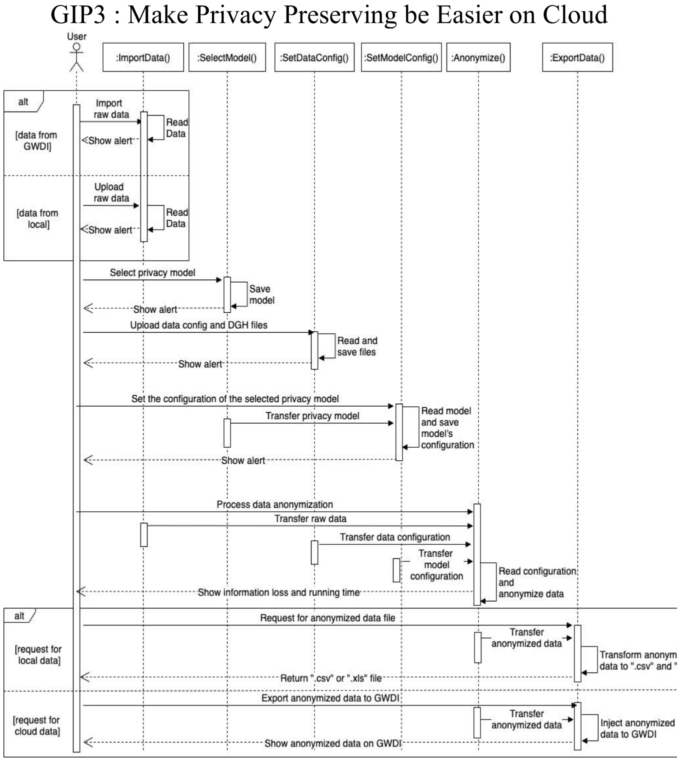


Fig. 5. UML sequence diagram of user's anonymization requests on GIP3

Table 2. The supported privacy preserving models & information loss metrics

Privacy Preserving Model	information loss metric			
	DM[2]	C_{avg} [9]	LM[19]	NCP[21]
K-anonymity[16]	C	C	C	C
L-diversity[10]	C		C	C
Anatomy[18]				
K^m-anonymity[17]			C	C
(K,K^m)-anonymity[13]				C

The system processing layer receives the request from the user presentation layer and reads in the raw data for pre-processing, and then the pre-processed data is anonymized by processing. After anonymization, the information loss is calculated and saved. Finally, information loss and anonymous data are sent back to the user presentation layer. Fig. 6 shows a UML activity diagram for the system processing layer to handle a privacy preserving task.

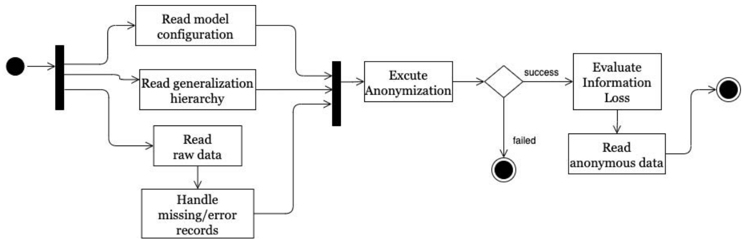


Fig. 6. UML activity diagram for the system processing layer to do a task

4. Experiment and Evaluation

We further verified the effectiveness of GIP3 in data privacy preserving by experiments. Given the widespread use of the classic method called k-anonymity, and the fact that GIP3 also supports k-anonymity, we compared it to UTD-AT[7] by experiment. Also, we noted that few tools implement km-anonymity so far, while it is an efficient and classic approach to handle transactional data, so we also demonstrated the feasibility of performing km-anonymity on GIP3.

4.1 Datasets and Experiments

Table 3 shows the datasets used in the experiments and the configurations of the experiments.

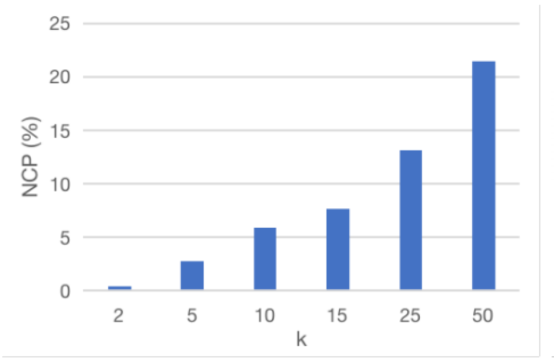
Table 3. Datasets and Experiments Configurations

Experiment Code	Exp.1	Exp.2	Exp.3
Dataset	Adult[1]	Informs[2]	IptvData
Records Count	32561	102579	62247
Privacy Preserving Model	k-anonymity	k-anonymity	k^m-anonymity
Information Loss Metric	NCP	NCP	NCP
Quasi Identifier Attributes	age, work class, marital status, class	DOBMM, marry, gender, RACEX, EDUCYEAR	DeviceID (Identifier)
Sensitive Attribute	occupation	income	ChannelNumber

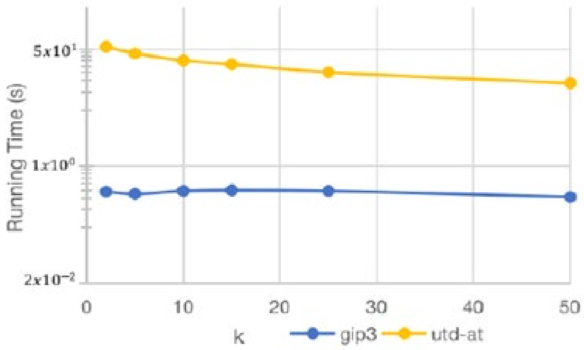
Since UTD-AT only supports command line operations, we used the anonymize server to process approach directly in the experiment. The environment is as follows: operating system: Ubuntu14.04.5LTS; CPU: Intel(R) Core(TM) i5-444; Memory: 8GB.

4.2 Results and Analysis

In Exp.1 and Exp.2, we selected k= {2,5,10,15,25,50}. Since UTD-AT does not provide NCP information loss metric, we only compared execution time. The results of Exp.1 are shown in Fig.7, and the results of Exp.2 are shown in Fig.8.

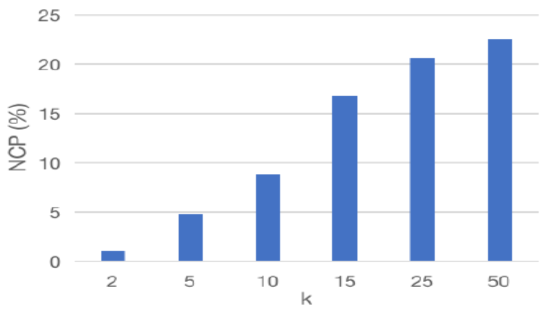


(a) Information loss NCP on GIP3

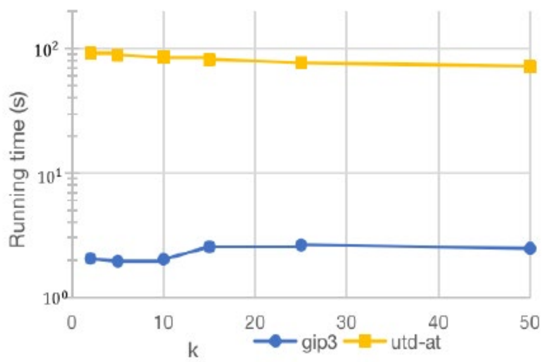


(b) Running time of GIP3 vs UTD-AT

Fig. 7. GIP3 vs UTD-AT for k-anonymity on Adult dataset



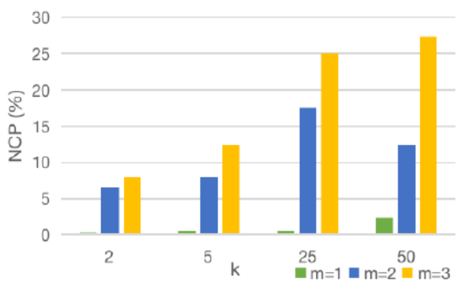
(a)Information loss NCP on GIP3



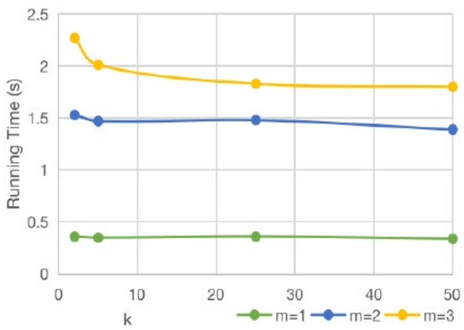
(b) Running time of GIP3 vs UTD-AT

Fig. 8. GIP3 vs UTD-AT for k-anonymity on Informs dataset

As shown in Fig. 7 snd Fig. 8, as k increases, information loss increases. This is because the number of records in equivalence classes increases, which requires more coarse-grained generalization. As k increases, the number of partition executions decreases, and the running time decreases. As the number of records in the dataset increases, the number of partition executions increases, resulting in longer execution times. While execution time is very important to the user, GIP3 is faster than UTD-AT, in which one of the important reasons is that UTD-AT maps all the category attributes to continuous numeric intervals and then continuing with the median partitioning method, while GIP3 uses the generalization hierarchy to handle category attributes directly. In addition, it is found that UTD-AT requires configuration of SQLite database, while , GIP3 provides SaaS cloud data management system for users, in which users do not need any local configuration.
In Exp.3, we selected k= 2,5,25,50 and m= 1,2,3 to conduct multiple experiments, in which the results are shown in Fig. 9.



(a) Information loss NCP on GIP3



(b) Running time of GIP3

Fig. 9. Processing k^m-anonymity on iptvdata dataset on GIP3

As shown in Fig. 9, both k and m are positively correlated with information loss, which is due to the decline of data utility caused by more coarse-grained generalization. In addition, as the m increases, the execution time increases significantly, which is due to the increasing number of generalization executions caused by the increase of Cm, where N represents the number of all values of transactional attribute. According to the result, when processing 60k data, the execution time of GIP3 is within 5 seconds, which can give users a satisfactory experience.

5. Conclusion

In this paper, we proposed a cloud data privacy preserving platform based on SaaS. This platform provides users with a convenient and efficient approach of data anonymization, and supports a variety of privacy preserving models and information loss metrics. It also provides users with data management services. It is a SaaS cloud data service platform that integrates data publishing, data statistics, data maintenance and data privacy preserving. We verified the avail- ability and efficiency of GIP3 via a series of experiments. In the future, GIP3 can be extended and optimized in the following aspects: (1)Providing the functionality API of GIP3, by which developers can call anonymization methods in their own system. (2)Providing risk assessment on GIP3, in which users can easily suppress tuples using it. (3)Providing more friendly graphical interface of generalization hierarchy for users to easily design.

References

[1] 1. Abdelhameed, S.A., Moussa, S.M., Khalifa, M.E.: Privacy-preserving tabular data publishing: a comprehensive evaluation from web to cloud. Computers & Security 72, 74–95 (2018)

[2] 2. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: 21st International conference on data engineering (ICDE'05). pp. 217–228. IEEE (2005)

[3] 3. Dai, C., Ghinita, G., Bertino, E., Byun, J.W., Li, N.: Tiamat: a tool for inter- active analysis of microdata anonymization techniques. Proceedings of the VLDB Endowment 2(2), 1618–1621 (2009)

[4] 4. Dandekar, A., Basu, D., Kister, T., Poh, G.S., Xu, J., Bressan, S.: Privacy as a service: publishing data and models. In: International Conference on Database Systems for Advanced Applications. pp. 557–561. Springer (2019)

[5] 5. Geetha, P., Naikodi, C., Setty, S.L.N.: Design of big data privacy framework—a balancing act. In: Advances in Data Sciences, Security and Applications, pp. 253– 265. Springer (2020)

[6] 6. Hundepool, A., De Wolf, P., Bakker, J., Reedijk, A., Franconi, L., Polettini, S., Capobianchi, A., Domingo-Ferrer, J.: mu-argus user's manual version 5.1. Statistics Netherlands: The Hague, The Netherlands (2018)

[7] 7. Kantarcioglu, M., Inan, A., Kuzu, M.: Utd anonymization toolbox (2012)

[8] 8. Ke, C., Huang, Z., Cheng, X.: Privacy disclosure checking method applied on collaboration interactions among saas services. IEEE Access 5, 15080–15092 (2017)

[9] 9. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k- anonymity. In: 22nd International conference on data engineering (ICDE'06). pp. 25–25. IEEE (2006)

[10] 10. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1), 3–es (2007)

[11] 11. Pacheco, V., Puttini, R.: Defining and implementing connection anonymity for saas web ser-vices. In: 2012 IEEE Fifth International Conference on Cloud Computing. pp. 479–486. IEEE (2012)

[12] 12. Poulis, G., Gkoulalas-Divanis, A., Loukides, G., Skiadopoulos, S., Tryfonopoulos, C.: Secreta: A tool for anonymizing relational, transaction and rt-datasets. In: Medical Data Pri-vacy Handbook, pp. 83–109. Springer (2015)

[13] 13. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: Anonymizing data with relational and transaction attributes. In: Joint European Conference on Machine Learn-ing and Knowledge Discovery in Databases. pp. 353–369. Springer (2013)

[14] 14. Prasser, F., Eicher, J., Spengler, H., Bild, R., Kuhn, K.A.: Flexible data anonymiza-tion using arx—current status and challenges ahead. Software: Practice and Experience (2020)

[15] 15. Preuveneers, D., Heyman, T., Berbers, Y., Joosen, W.: Systematic scalability assessment for feature oriented multi-tenant services. Journal of Systems and Soft- ware 116, 162–176 (2016)

[16] 16. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncer-tainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)

[17] 17. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set- valued data. Proceedings of the VLDB Endowment 1(1), 115–125 (2008)

[18] 18. Xiao, X., Tao, Y.: Anatomy: Simple and effective privacy preservation. In: Proceed- ings of the 32nd international conference on Very large data bases. pp. 139–150 (2006)

[19] 19. Xiao, X., Tao, Y.: Personalized privacy preservation. In: Proceedings of the 2006 ACM SIGMOD international conference on Management of data. pp. 229–240 (2006)

[20] 20. Xiao, X., Wang, G., Gehrke, J.: Interactive anonymization of sensitive data. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. pp. 1051–1054 (2009)

[21] 21. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.C.: Utility-based anonymiza- tion us-ing local recoding. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 785–790 (2006)

[22] 22. Yang, C., Pan, S., Li, R., Liu, Y., Peng, L.: A coding-free software framework for de-veloping lightweight web data management systems. Applied Sciences 10(3), 865 (2020)

A Demonstration

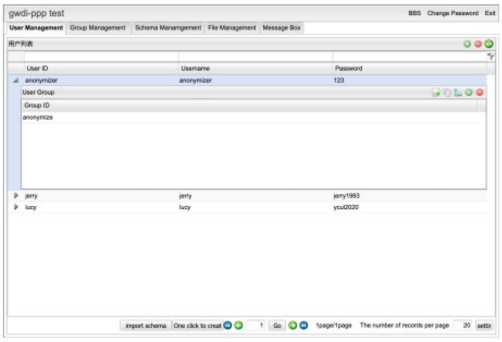
For the convenience of using GIP3 in detail, we demonstrate a specific case of a user requesting for data privacy preserving in this appendix.
First, a tenant user should register a tenant account on GIP3 for constructing his/her tenant data management

system, then create the accounts for the users of constructed system in GIP3 and inform them. Next, the user of a tenant can log in the constructed system, manage and anonymize data by k-anonymity approach in the system. Following us, you can learn more and process your data anonymization.

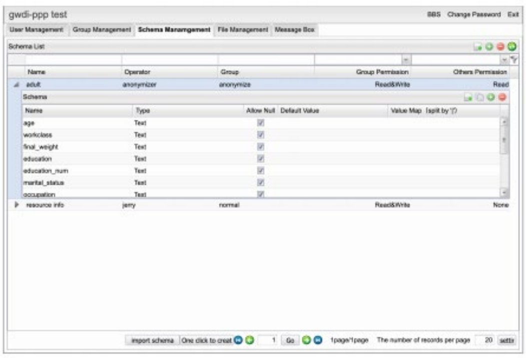
– Step1. We first visit GIP3's website and register a tenant account named tom2020, and then we log into data management system as the tenant tom2020. Then, we create a user named anonymizer and add this user to the group named anonymize, and we create a data schema named adult with its fields, as shown in Fig. 10.

GIP3 : Make Privacy Preserving be Easier on Cloud

(a) Register a tenant account on GIP3



(b) Create the user named anonymizer and add this user to group anonymize



(c) Create the schema named adult with elds

Fig. 10. Tenant creates and configures system on GIP3

Step2. We log into GIP3 as the user anonymizer, and then upload data in adult schema, in which anonymizer has the read/write permissions on adult schema. After uploading, data is shown in the system as shown in Fig. 11.

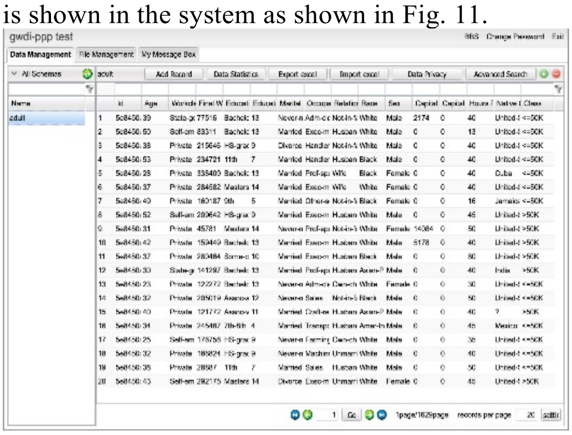


Fig. 11. User anonymizer uploads data to GIP3

Step3. We click the "data privacy" button to get privacy preserving service. And first, we click the button "import" to submit current schema's data as raw data, as shown in Fig. 12.

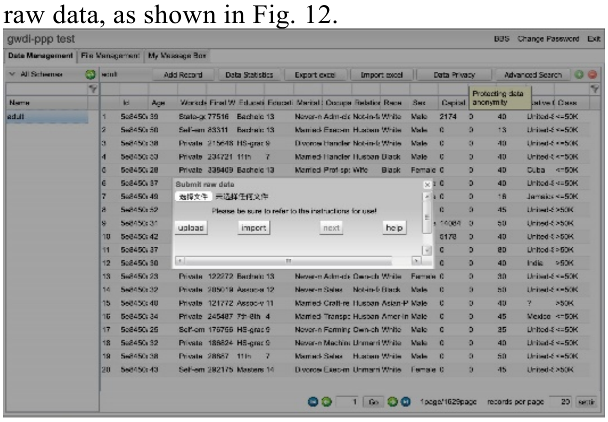


Fig. 10. User submits raw data

Step4. After submitting raw data, we click "next" button to select **privacy** preserving model. As shown in Fig. 13, there are five different models for user to select, in which we select k-anonymity to use.

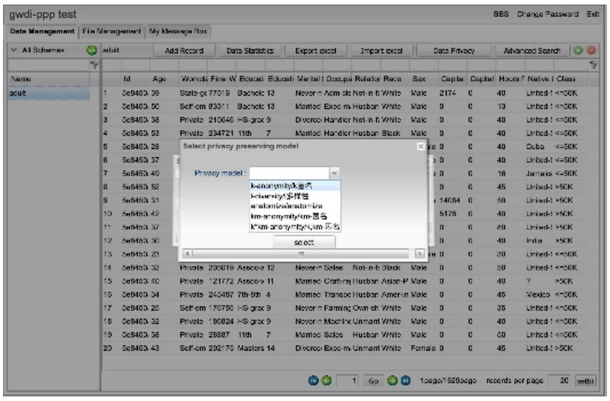


Fig. 11. User selects privacy preserving model

Step5. After selecting model, we should upload our configuration files. In thisstep, we define data's quasi identifier attributes(QIA), QIA's type(category or numeric), QIA's generalization hierarchy and sensitive attribute. We pack- age all the configuration files as zip and upload the files, as shown in Fig.14.

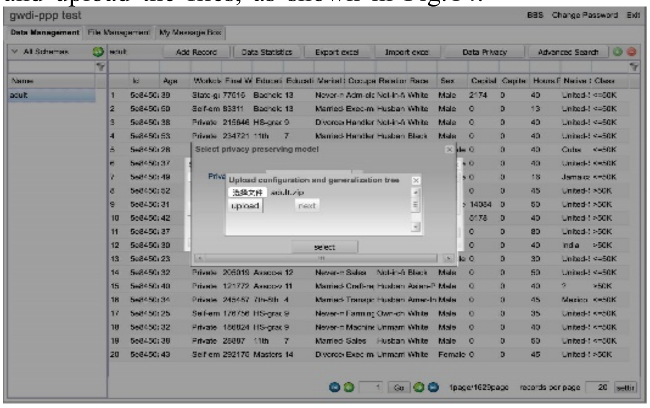
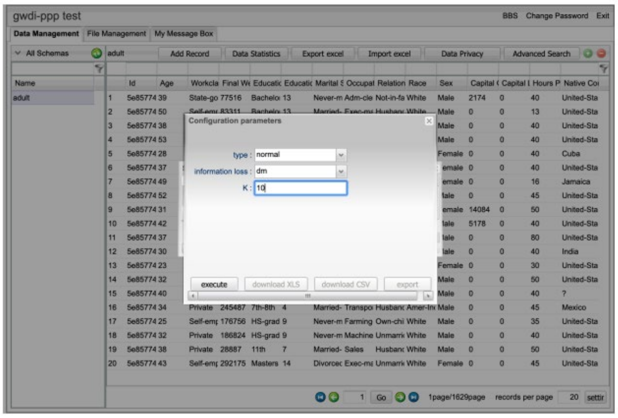


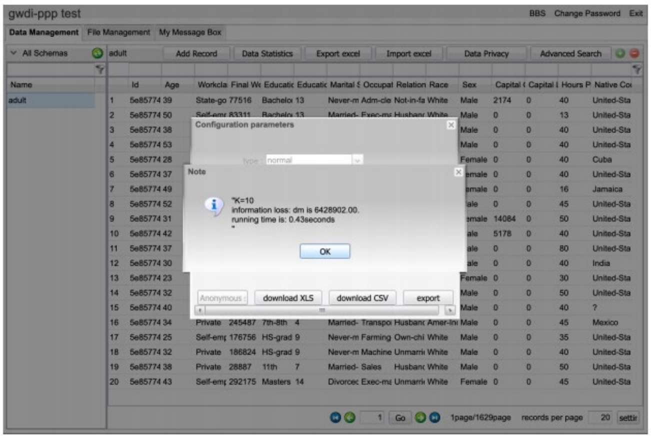
Fig. 12. User uploads related configurations

Step6. After uploading configurations of data, we should set parameters of models. In step4, we selected k-anonymity as the privacy preserving model, so we set parameters of k-anonymity in this step. First, we select normal type as performing type which means we should set the value of k manually. Next, we select DM as in-formation loss metric to evaluate utility of data. Finally, we set the value of k = 10, as shown in Fig. 15(a). Then we request for executing anonymization, when anonymization is successful, it returns information loss and execution time to user, as shown in Fig. 15(b).

GIP3 : Make Privacy Preserving be Easier on Cloud

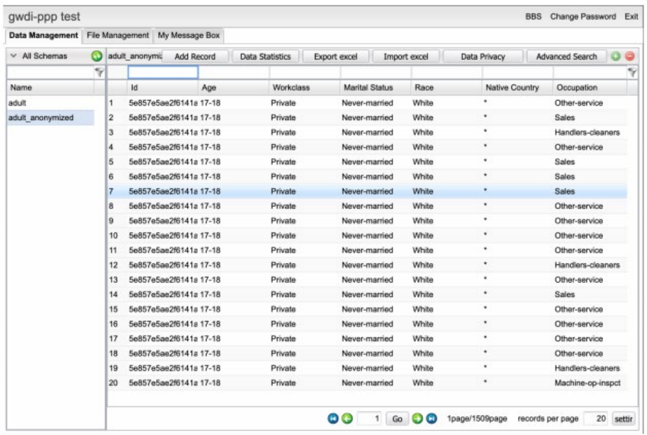


(a) User sets privacy preserving model's parameters

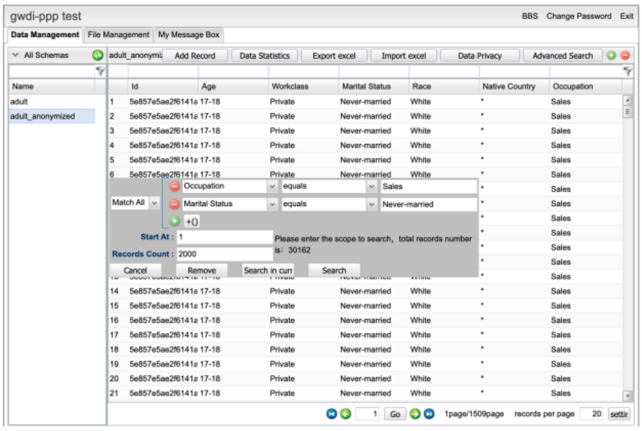


(b) Information loss and running time of anonymization

Step7. In this step, we request for the data after anonymization. We click the "ex-port" button to export anonymous data to the cloud data management system. After exporting data and refreshing schemas list, the anonymous data is shown in the system which we can manage and analyze data in it, as shown in Fig. 16.



(a) Data after anonymization on GIP3



(b) Query in the first 2000 records where Occupation "Sales" and Marital Status "Never married"

The above is an example of using GIP3 for data privacy preserving. We hopeit to be helpful for the users interested in GIP3 to deal with data privacy presering. More detailed information can be found in the website of http://www.iaihust.com, and the other website of http://www.web3.org.cn:8080/top1 can also be accessed via China Education and Research Network (CERNET).