Hindawi
Security and Communication Networks
Volume 2023, Article ID 9795620, 1 page
https://doi.org/10.1155/2023/9795620

Retraction

Retracted: An Automated Data Desensitisation System Based on the Middle Platform

Security and Communication Networks

Received 5 December 2023; Accepted 5 December 2023; Published 6 December 2023

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Xu and H. Zhou, "An Automated Data Desensitisation System Based on the Middle Platform," *Security and Communication Networks*, vol. 2022, Article ID 6888441, 8 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 6888441, 8 pages
https://doi.org/10.1155/2022/6888441

Research Article

An Automated Data Desensitisation System Based on the Middle Platform

Lei Xu and Haocheng Zhou

Jiangsu Electric Power Information Technology Co., LTD, Nanjing, Jiangsu 210000, China

Correspondence should be addressed to Lei Xu; 2010651103@hbut.edu.cn

Received 29 July 2022; Revised 2 September 2022; Accepted 7 September 2022; Published 21 September 2022

Academic Editor: C. Venkatesan

Built on top of a big data platform, the Middle Platform develops data through abstraction, sharing, and reuse capabilities to provide data products and data services for upper-level business development. While fully analysing and mining the intrinsic value of data, privacy and sensitive information in the data must also be protected, so the Middle Platform needs a data desensitisation system to ensure the safe and open use of data. In order to solve the problems of high usage costs, low efficiency, and lack of standardised results of desensitisation that exist in conventional data desensitisation systems, an automated desensitisation system with data assets, access control, and desensitisation strategies as the main modules is established using an adaptive method of generating dynamic desensitisation rules, combined with a security monitoring mechanism of sensitivity classification and two-level permissions. The system optimises the configuration structure to obtain stable and reliable desensitisation results and efficiently respond to diverse business needs. Users are able to get rid of complex rule management and focus on the data usage itself.

1. Introduction

The Middle Platform is a data service product featuring data aggregation and governance of cross-domain data (data from different data sources), which drives business development with data development and, thus, improves overall development efficiency. When large amounts of data are deposited in the Middle Platform, accessed, queried, processed, and calculated with high frequency, the risk of compromising user privacy or trade secrets contained in the data grows at a high rate. As a result, international regulations in various industries require data to be privacy-protected before they can be made available for use [1].

In current privacy protection practices, data desensitisation is a common technical means of dealing with high efficiency. The original sensitive data is transformed into less sensitive desensitised data by applying a series of data distortions to the more sensitive raw data. Although the desensitised data is distorted to a certain extent, it still has some data value and this distortion is acceptable when balanced against data security and usability. In data desensitisation technology, a desensitisation algorithm is a method of data distortion used in the desensitisation process, which is applied to specific sensitive data to form a desensitisation rule. Desensitisation rules are named after sensitive data and there can be multiple desensitisation rules for one type of sensitive data.

The data desensitisation system in the Middle Platform needs to develop a set of suitable desensitisation rules for a sensitive data set according to the user's business requirements and convert the sensitive data into desensitised data for output. Conventional desensitisation systems have a number of built-in desensitisation rules for each sensitive data item, and the desensitisation task is performed by manually configuring the desensitisation rules for each desensitised data item. This system is essentially management of rules, making it necessary for the user to learn the various desensitisation algorithms and rules before the rules can be configured. Not only is a lot of learning and operational costs invested, but the more sensitive the data and the

more complex the business requirements, the more significant the reduction in efficiency; the degree of manual influence is also too deep, and the output desensitised data are not standardised enough to serve the business requirements consistently and even less able to cope with the dynamic desensitisation requirements with high real-time requirements.

To solve the abovementioned problems, this study discusses the desensitisation system in a comprehensive manner from multiple perspectives of data management, desensitisation rules, and application scenarios with the background of scenarios and data in the power industry, designs a desensitisation strategy based on adaptive theory, and manages the system user roles with a two-tier mechanism combining sensitive permissions and business requirements. An automated desensitisation system consisting of data assets, access control, and desensitisation strategies as the main modules, with the generation of dynamic desensitisation rules as the core, has been established. The system enables rapid batch desensitisation by configuring desensitisation strategies while taking into account diverse desensitisation needs. The desensitisation results are stable and reliable, and more desensitisation rules can be evolved by adding desensitisation algorithms and desensitisation strategies, making it easier to expand business requirements.

2. Materials and Methods

The automated desensitisation system based on the Middle Platform is shown in Figure 1.

On the basis of the data assets, the desensitisation system then obtains a collection of sensitive data based on access control, sets the conditions and forms of data opening, and formulates a desensitisation strategy according to the abstracted business requirements. Ultimately, the desensitisation strategy matches the security and availability requirements with the desensitisation strength and algorithm weights, respectively, generating dynamic desensitisation rules to perform the desensitisation task.

2.1. Data Assets. The Middle Platform aggregates data from multiple data sources, sorts out the various data structures, and plans the value of the data from a business perspective so that the data forms data assets. For desensitised systems, the management of data assets mainly includes sensitive data identification, sensitive data classification, and sensitive data classification.

2.1.1. Identifying Sensitive Data. Sensitive data usually have a specific or agreed encoding format and rules, and the system can use matching algorithms such as regular expressions and keywords to capture the characteristic fields and obtain sensitive data sets. According to the standards related to the protection of sensitive information in the power industry [2]. Table 1 lists some common sensitive data and their encoding characteristics.

2.1.2. Classifying Sensitive Data. The classification can be based on the source, content and use of the data, etc. and is usually set by the business unit. In this system, the data classification determines the type of business of the data user and is related to the user role setting of the system. Common sensitive data in the power industry is divided into three broad categories [3]: production data, marketing data, and management data. The data listed in Table 1 belong to production data and is mostly used for business services such as development, testing, querying, and sharing; marketing data and management data are mostly used for querying business services.

2.1.3. Grade for Sensitivity. The classification is based on the impact of a breach of the security attributes of the data and is usually set by the business unit. Common sensitive data in the power industry is classified into four levels of confidentiality, with 1 to 4 being progressively more sensitive [4].

In this system, there are two classifications of sensitive data throughout its life cycle.

Definition 1. The classification of the sensitivity of the original data are called the original confidentiality level. It is denoted by LO.

Definition 2. The sensitivity level of desensitised data are called the desensitised confidentiality level. It is denoted by LF.

The sensitive data in Table 1 are divided into their original classification to obtain Table 2.

After the raw data have been deformed, the desensitised data are less sensitive and more secure. The change in sensitivity from raw to desensitised data depends mainly on the desensitisation intensity [5]. The desensitisation intensity is defined as 3 to 1 from strong to weak according to Table 3, respectively.

2.2. Access Control. Access control refers to the configuration of "roles" and their attributes for users of the desensitisation system from a two-dimensional perspective of business applications and user confidentiality, defining who is to be desensitised and what is to be desensitised. The "role" describes the user's requirements for security and availability of the desensitised results.

Definition 3. The level of sensitivity at which a system allows a user to use data are called sensitive permission. It is denoted by LA, and expresses the security requirements. Again there are 4 levels, with progressively higher permissions from 1 to 4.

For each type of sensitive data, the system must perform desensitisation when the user's $LA < LO$ and need not perform desensitisation when the user's $LA \geq LO$.

Definition 4. The matrix of data attributes that the system abstracts and quantitatively assigns to business applications is called business permissions. It is denoted by RA, and expresses usability requirements. It consists of three data

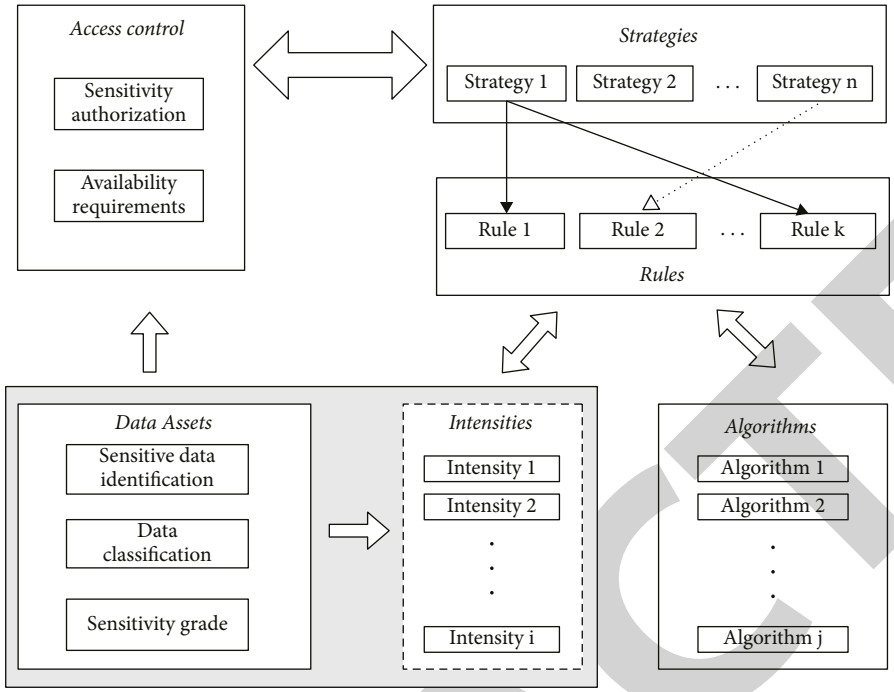

Figure 1: Middle Platform desensitisation system architecture.

Table 1: Encoding structures and rules of common electric power sensitive data.

Sensitive data item	Encoding structure and rule	Example
Client's no.	10-digit serial number	0257349261
ID card	18 digits, 6-digit area code + 8-digit birth date + 4-digit serial number	330101197701014237
Bank card no.	13–19 digits, issuing bank number + card type number + serial number	9558801202106562334
Electricity address	City + district/county + street/town + community/village + road + house number	16th floor, no. 56, Huaqiao road, Gulou district, Nanjing
Mobile no.	11 digits, 3-digit network identifier + 4-digit area code + 4-digit serial number	13088886666
Electricity consumption	Random number	250, 374, 499
Settlement date	8 digits, 4-digit "year" + 2-digit "month" + 2-digit "day"	20210101

Table 2: Original confidentiality levels of common electric power sensitive data.

Sensitive data item	Original confidentiality level *LO*
Client's no.	1
ID card	4
Bank card no.	4
Electricity address	3
Mobile no.	3
Electricity consumption	2
Settlement date	1

attributes: integrity, reality, and repeatability [6]. Assigning a value of 0 or 1 to each attribute yields Table 4.

The main data openness scenarios are development, testing, query, and sharing operations [7]. Therefore, four roles are set up in this system. Taking the use of production data as an example, the specific configuration items are shown in Table 5.

Table 3: Relationship between *LO*, *LF*, and desensitisation intensity.

LO	*LF*		Desensitisation intensity
4	1	3	High
	2	2	Medium
	3	1	Low
3	1	2	Medium
	2	1	Low
2	1	1	Low
1	1	0	No need for desensitisation

Table 4: Description of desensitisation results availability requirements.

Attribute	Algorithm description	1	0
Integriality	Whether to keep the encoding structure intact	Y	N
Reality	Whether to reflect data real semantics	Y	N
Repeatability	Are data distortion parameters controllable?	Random Keyless	Quantitative Keyed

Table 5: Desensitisation system role configuration.

Role	Business scope	Sensitive permission **LA**	Business permissions **RA**		
			Integriality	Reality	Repeatability
Developer	Development	3	1	0, 1	1
Tester	Test	2	1	1	0, 1
Administrator	Enquiry	4	—	—	—
Operator	Share	1	1	0	0

— LA ≥ LO, there is no need for desensitisation.

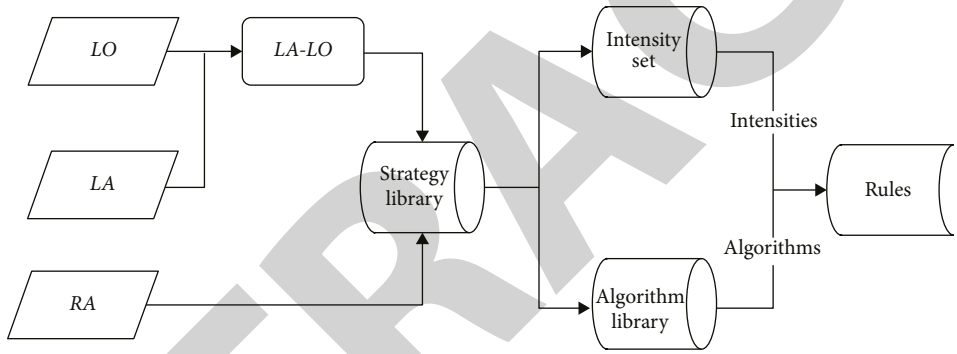

Figure 2: Dynamic rule model for desensitisation system.

2.3. Desensitisation Strategy. Desensitisation algorithms are data deformation methods used in the desensitisation process, and the application of desensitisation algorithms to specific sensitive data results in desensitisation rules [8]. The system introduces a desensitisation strategy that establishes a link between the data usage requirements of the user's business and the library of desensitisation algorithms. Using an adaptive strategy model, the desensitisation strategy is designed using the desensitisation strength and the desensitisation algorithm (weight) as factors, while breaking the limits of fixed desensitisation rules, and using the method of generating dynamic desensitisation rules (Figure 2) to perform data desensitisation tasks in various application scenarios.

We consider a sensitive dataset $D = \{D_1, D_2 \ldots D_k\}$, containing k kinds of sensitive data; then,

The original confidentiality level of the data $LO = \{LO_1, LO_2 \ldots LO_k\}$;

The desensitised confidentiality level of the data $LF = \{LF_1, LF_2 \ldots LF_k\}$;

The sensitive permission of the user $LA = \{LA_1, LA_2 \ldots LA_k\}$.

2.3.1. Determining the Desensitisation Intensity Range. The user's sensitivity level must be higher than the final desensitisation level of the data in order for the data to be secure for open use. That is, $LA_i \geq LF_i$;

Via desensitisation, intensity $I_i = LO_i - LF_i$, $i = 1, 2 \ldots k$.

Then, $I_i \geq LO_i - LA_i$, and I_i takes a value among 0, 1, 2, and 3.

2.3.2. Assigning Desensitisation Intensity and Desensitisation Algorithms. When the abstract expression of the user's business requirement is RA, the system matches the desensitisation strategy as DS.

The algorithm weight W_i for each desensitised data is obtained from the strategy analysis and combined with the intensity range obtained in the previous step; the system selects the final algorithm A_i and intensity I_i in the desensitisation algorithm library and desensitisation intensity grading table.

2.3.3. Generating Dynamic Desensitisation Rules. At the strength of I_i, the data deformation is performed with an algorithm A_i, which constitutes a desensitisation rule R_i for

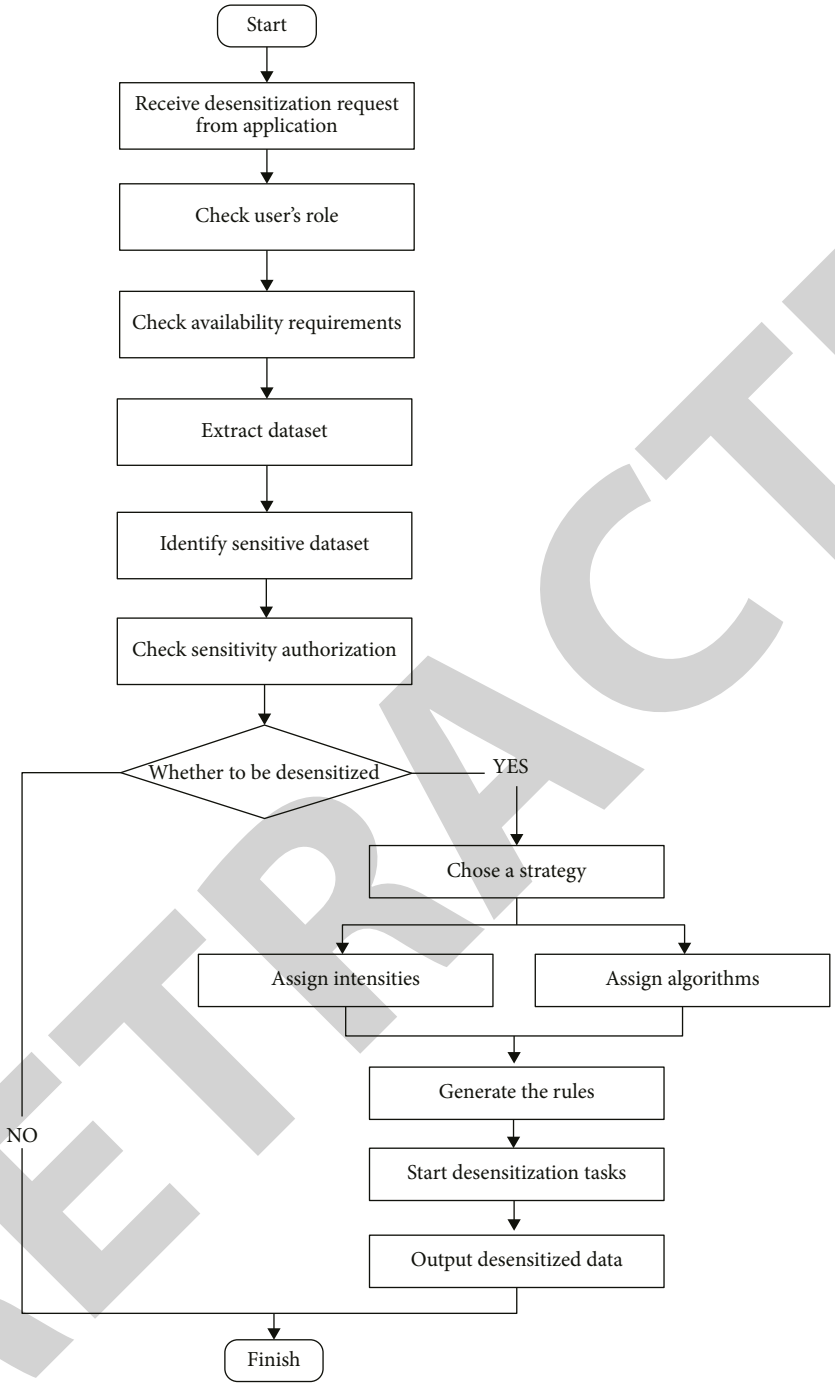

Figure 3: Automated data desensitisation flow chart.

sensitive data i, then there is a desensitisation rule set $R = \{R_1, R_2 \ldots R_k\}$ for the entire set of sensitive data.

3. Results

It is assumed that the application data are tested on the Middle Platform of STATE GRID JIANGSU ELECTRIC POWER COMPANY data for the bill payment function.

The system performs desensitisation tasks according to the flow in Figure 3.

The user's role for this data use application is "Tester."

A production data set is extracted from the Middle Platform, and the data to be desensitised includes: customer number, telephone number, bank card number, electricity consumption address, electricity consumption, and billing date. The abovementioned sensitive information can be identified using regular expressions and keyword algorithms to obtain the sensitive data set. Combining Tables 1 and 2,

Table 6: Encoding structure and original classification of sensitive data.

No.	Sensitive data item	Encoding structure and rule	Example	*LO*
1	Client's no	10-digit serial number	0257349261	1
2	Mobile no.	11 digits, 3-digit network identifier + 4-digit area code + 4-digit serial number	13088886666	3
3	Bank card no.	13–19 digits, issuing bank number + card type number + serial number	9558801202106562334	4
4	Electricity address	City + district/county + street/town + community/village + road + house number	16th floor, no. 56, Huaqiao road, Gulou district, Nanjing	3
5	Electricity consumption	Random number	250, 374, 499	2
6	Settlement date	8 digits, 4-digit "year" + 2-digit "month" + 2-digit "day"	20210101	1

Table 7: Desensitisation algorithms.

No.	Sensitive data item	Desensitisation intensity values
1	Client's no	0
2	Mobile no.	1, 2
3	Bank card no.	1, 2, 3
4	Electricity address	1, 2
5	Electricity consumption	0
6	Settlement date	0

Table 8: Desensitisation algorithms and weights.

Algorithm	Description	Example	Weight
Mask	Use symbol "*" to replace parts of the data, with the data length unchanged.	13088886666 -> 130********	1, 0, 1
Floor	Take an integer	—	0, 0, 0*
Hashing	Map data into a fixed-length string	13088886666 -> abcdef	0, 0, 1
Truncation	Cut parts of the data	13088886666 -> 130	0, 0, 1
Shift	Add an constant offset	13088886666 -> 13088886670	1, 1, 0
Synthesis	Simulate new data to replace the original data	13088886666 -> 13011007788	1, 1, 1
Rearrange	Sort a column of values upside-down	—	0, 0, 0*

Table 6 shows the encoding rules and the original confidentiality level *LO* for these sensitive data.

According to Table 5, the "Tester's" sensitive permission LA = 2 and business permissions of the test application $RA = [1, 1, 1]$ or $[1, 0, 1]$.

Then the desensitisation intensity set $I = \{I_1, I_2, I_3, I_4, I_5, I_6\}$ for the 6 sensitive data, I_i takes the values in Table 7.

A library of desensitisation algorithms is available in the system, as shown in Table 8 [9].

The various sensitive data in Table 6 were graded for desensitisation intensity to get Table 9 [10].

The desensitisation strategies available in the system are shown in Table 10.

If the user specifies the strategy "minimum enough," the desensitisation strengths and desensitisation algorithms for the 6 sensitive data types can be obtained by filtering from the strength classification Table 8 and the algorithm weight vector Table 9, and the composed desensitisation rules and final desensitisation results are shown in Table 11.

4. Discussion

The desensitisation system differs from traditional systems in, that it, transforms the management of desensitisation rules into the management of desensitisation strategies, making the desensitisation strategies more strongly coupled with the business applications, and freeing the desensitisation rules to focus on data deformation processing. The sensitivity classification of data and the abstraction of application scenarios are used as a means to quantitatively assess the security and availability requirements of business applications for desensitisation results, enabling the system to achieve standardised and automated desensitisation. The criteria for sensitivity classification and application scenario abstraction are set by the business unit and may be dynamically adjusted as the data changes in terms of aggregation, volume, and scale, and as business needs expand in terms of diversity and timeliness. This is a future research direction for this system. However, as both the desensitisation intensity table and the algorithm weighting table are only relevant to the data deformation process itself, they are independent of the grading scale and

Table 9: Sensitive data desensitisation intensity grading.

No.	Sensitive data item	*LO*	Desensitisation intensity grade	Reserved bit	Desensitised bits	Example
1	Client's no	1	0	Keep all the 10	0	0257886496 -> 0257886496
2	Mobile no.	3	2	Keep the top 3	The bottom 8	13088886666 -> 130********
			1	Keep the top 3 & bottom 4	The middle 4	13088886666 -> 130****6666
3	Bank card no.	4	3	Keep the top 4	The bottom 15	9558801202106562334 -> 9558***************
			2	Keep the top 4 & bottom 4	The middle 11	9558801202106562334 -> 9558**********2334
			1	Keep the top 8 & bottom 4	The middle 7	9558801202106562334 -> 95588012*******2334
4	Electricity address	3	2	Keep "city" & "district/county" & "house no."	Street/town + community/village + road	16th floor, no. 56, Huaqiao road, Gulou district, Nanjing -> 16th floor, no. 56, *** road, Gulou district, ***
			1	Keep "street/town" & "community/village" & "road"	City + district/county + house no.	16th floor, no. 56, Huaqiao road, Gulou district, Nanjing -> ***, ***, Huaqiao road, Gulou district, Nanjing
5	Electricity consumption	2	1	[1]	[1]	[1]
6	Settlement date	1	0	Keep all the 8	0	20210108 -> 20210108

[1]For random numerical type sensitive data, because there was no coding structure limit, the deformation effects were mainly affected by the algorithm. So, the default desensitisation intensity was equal to the configured value.

Table 10: Desensitisation strategies of the system.

Strategy name	Description	Intensity	Algorithm weighting
Maximum strength	Take the highest value of desensitisation intensity	Max $\{I_i\}$	*RA*
Same algorithm	Consistent algorithms for all sensitive data	I_i	configuration item, such as "mask"
Minimum enough	Take the lowest value of desensitisation intensity	Min $\{I_i\}$	*RA*
Best simulation effect	Data emulation as far as possible on a confidential level permit	I_i	*RA*

Table 11: Desensitisation rules and results under "minimum enough."

Rule	Intensity	Desensitised bits	Algorithm	Result
Client's no.	0	—	—	0257349261 -> 0257349261
Mobile no.	1	The middle 4	Mask	13088886666 -> 130****6666
Bank card no.	1	The middle 7	Synthesis	9558801202106562334 -> 9558801223645452334
Electricity address	1	City + district/county + house no.	Synthesis	16th floor, no. 56, Huaqiao road, Gulou district, Nanjing -> 1th floor, no. 1 Huaqiao road, Gulou district, Nanjing
Electricity consumption	0	—	—	250, 374, 499 -> 250, 374, 499
Settlement date	0	—	—	20210101 -> 20210101

business abstraction methods, so they will not be affected by changes in the standards set by the business sector, and the system remains well suited to the open use of sensitive data in the Middle Platform.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Conceptualization, methodology, formal analysis, resource collection, original draft preparation, supervision, project administration, and funding acquisition were done by Xu Lei. Validation, investigation, data curation, and review and editing were done by Haocheng Zhou. Both authors have read and agreed to the published version of the manuscript.

Acknowledgments

This research was funded by the Research and Application of Data Security Protection Technology Based on Middle Platform, under grant no. 5210ED2100L. The APC was funded under 5210ED2100L.

References

[1] G. Market, "Guide for Data Masking," 2019, https://www.gartner.com/en/documents/4009400.

[2] S. Ye, "Desensitization evaluation and system iplementation for big data on electric power," *Heilongjiang Electric power*, vol. 42, no. 4, pp. 366–371, 2020.

[3] China Electricity Council Standardization Management Center, *Implementation Specification for Power Data Masking*, China Electricity Council Standardization Management Center, Beijing, China, 2020.

[4] China Southern Power Grid Company Limited, *Open Data Asset Classification and Classification Implementation Guide*, China Southern Power Grid Company Limited, Shenzhen, China, 2021.

[5] TencentCloud, "TencentCloud," 2021, https://cloud.tencent.com/developer/article/1636078.

[6] M. Riesselman and C. Miettinen, "An Adaptive Desensitization Method and System for Sensitive Data," *The Journal of Immunology*, vol. 179, no. 4, pp. 2520–2531, 2007.

[7] CSDN, "CSDN," 2020, https://blog.csdn.net/meichuangkeji/article/details/107716491.

[8] J. Wang, M. Xu, and K. Lu, "The research of adaptive data desensitization method based on Middle platform," *Applied Mathematics and Nonlinerar Sciences*, vol. 2022, Article ID 5348637, 7 pages, 2022.

[9] Cyberspace Administration of China, *Measures for Data Security Management*, Cyberspace Administration of China, Beijing, China, 2019.

[10] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," *Journal of the American Medical Informatics Association*, vol. 17, no. 2, pp. 169–177, 2010.