

EFFECTIVE DATA GOVERNANCE MODEL FOR
SAFEGUARDING FINANCIAL SENSITIVE
INFORMATION

ZHANG YUHE

UNIVERSITI KEBANGSAAN MALAYSIA

EFFECTIVE DATA GOVERNANCE MODEL FOR SAFEGUARDING
FINANCIAL SENSITIVE INFORMATION

ZHANG YUHE

PROJECT SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF
MASTER IN CYBER SECURITY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

MODEL TADBIR URUS DATA YANG BERKESAN UNTUK MELINDUNGI
MAKLUMAT SENSITIF KEWANGAN

ZHANG YUHE

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT UNTUK MEMPEROLEH IJAZAH SARJANA SIBER
KESELAMATAN

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI
2025

DECLARATION

I hereby declare that the work in this project is my own except for quotations and summaries which have been duly acknowledged.

I have not used any AI tools or technologies to prepare this report.

07 January 2025

ZHANG YUHE
P133718

ACKNOWLEDGEMENT

I am very fortunate to have Dr Hanis as my research supervisor, and I am very grateful to Dr Hanis for her advice and guidance. She has very rich knowledge of network security, and when I encountered difficulties in the research, she helped me provide a good idea to solve the problem, so that I learned a lot in the research.

At the same time, I would like to thank the Faculty of Science and Technology for their assistance and support. The faculty provides a good platform for my thesis research work, so that I can focus more on my research topic.

I would like to thank all the cyber security graduate students for their help and friendship during my studies at Universiti Kebangsaan Malaysia. When I encounter difficulties in study, they give me great encouragement and motivation and help me overcome great difficulties.

I would also like to thank the information security experts who participated in the evaluation and validation of the models in this article. They provided great support and help in the model verification of this study, and put forward good suggestions to help me better improve the research content of this study.

In addition, I would like to thank all the experts and teachers who provided valuable suggestions for this study. I hope that I can continue to study cyber security related content in the future and contribute to cyber security research.

ABSTRAK

Risiko kebocoran data adalah salah satu serangan keselamatan maklumat penting yang dihadapi oleh bidang kewangan, yang bukan sahaja akan menyebabkan kerugian ekonomi yang serius kepada perusahaan kewangan, tetapi juga merosakkan reputasi perusahaan dan kepercayaan pelanggan secara serius. Oleh itu, membina model tadbir urus keselamatan data yang boleh mempertahankan secara berkesan daripada risiko pelanggaran data telah menjadi kebimbangan utama. Model tadbir urus keselamatan data sedia ada terutamanya mempunyai beberapa masalah, seperti kekurangan pengenalpastian had model yang berkesan, keberkesanan terhad pencegahan risiko kebocoran data dan kekurangan mekanisme pengesanan model bersatu. Masalah ini menjadikan model tadbir urus keselamatan data sedia ada mempunyai beberapa had, tetapi juga meningkatkan risiko perusahaan dalam pengurusan peristiwa pelanggaran keselamatan data. Tujuan kajian ini adalah untuk membina model tadbir urus keselamatan data yang praktikal dan berkesan melalui perbincangan mendalam tentang had model tadbir urus data sedia ada, digabungkan dengan pengenalpastian risiko kebocoran data sebenar, supaya dapat mengurangkan risiko kebocoran data kewangan. perusahaan. Model yang dicadangkan terutamanya berdasarkan kajian kes dan kaedah pembinaan model, dan melalui tinjauan pakar keselamatan maklumat, keberkesanan model yang dicadangkan terhadap risiko kebocoran data disahkan. Ini akan menyediakan penyelesaian rangka kerja untuk tadbir urus keselamatan data dalam industri kewangan, dan membantu perusahaan kewangan dengan ketara mengurangkan risiko keselamatan pelanggaran data. Ia akan membantu perusahaan kewangan untuk memanfaatkan lagi nilai data, meningkatkan pengaruh perusahaan dan membantu meningkatkan faedah ekonomi perusahaan.

ABSTRACT

The risk of data leakage is one of the important information security attacks faced by the financial field, which will not only cause serious economic losses to financial enterprises, but also seriously damage the reputation of enterprises and the trust of customers. Therefore, building a data security governance model that can effectively defend against the risk of data breaches has become a major concern. The existing data security governance models mainly have some problems, such as lack of effective model limitation identification, limited effectiveness of data leakage risk prevention and lack of unified model validation mechanism. These problems make the existing data security governance model have some limitations, but also increase the risk of enterprise in data security breach event management. The purpose of this study is to build a practical and effective data security governance model through in-depth discussion of the existing data governance model limitation, combined with the identification of actual data leakage risk, so as to reduce the risk of data leakage of financial enterprises. The proposed model is mainly based on case study and model building methods, and through the survey of information security experts, the effectiveness of the proposed model against the risk of data leakage is verified. This will provide a framework solution for data security governance in the financial industry, and better help financial enterprises significantly reduce the security risk of data breaches. It will help financial enterprises to further tap the value of data, increase the influence of enterprises, and help improve the economic benefits of enterprises.

TABLE OF CONTENTS

		Page
DECLARATION		iii
ACKNOWLEDGEMENT		iv
ABSTRAK		v
ABSTRACT		vi
TABLE OF CONTENTS		vii
LIST OF TABLES		ix
LIST OF ILLUSTRATIONS		x
CHAPTER I	INTRODUCTION	
1.1	Research Background	1
1.2	Problem Statement	3
1.3	Research Questions	5
1.4	Research Objectives	6
1.5	Research Methodology	7
1.6	Research Scope	8
CHAPTER II	LITERATURE REVIEW	
2.1	Introduction	9
2.2	The Classification of Sensitive Data in the Financial Sector	10
	2.2.1 Data security standards	10
	2.2.2 Definition of sensitive data	13
	2.2.3 Methods of sensitive data identification	16
2.3	Comparison of Existing Sensitive Data Governance Models	19
2.4	Limitation of Existing Sensitive Data Models	25
	2.4.1 Comparative analysis of model design concepts	25
	2.4.2 A comparative analysis of the overall framework	26
	2.4.3 Analysis of model characteristics	27
CHAPTER III	METHODOLOGY	
3.1	Introduction	39

3.2	Research Methodology	39
3.3	The Proposed Model Based on Case Study	40
	3.3.1 Design concept	41
	3.3.2 Model comparative analysis	42
	3.3.3 Effectiveness evaluation	42
	3.3.4 Technical architecture analysis	42
	3.3.5 Compliance requirements	43
	3.3.6 Risk management	43
	3.3.7 Data life cycle management	43
	3.3.8 The proposed model based on the existing models	46
3.4	The Validation of Model Based on Survey	48
	3.4.1 The objective of model validation	49
	3.4.2 Contents	50
	3.4.3 Questionnaire information	52
	3.4.4 The result collection and analysis	54
	3.4.5 Expert interview	54
	3.4.6 Overview	54
CHAPTER IV RESULTS AND DISCUSSION		
4.1	The Proposed Data Security Governance Model	56
4.2	Validation of Proposed Model	78
	4.2.1 Introduction	78
	4.2.2 Results	78
	4.2.3 Analysis of result	90
CHAPTER V CONCLUSION AND FUTURE WORKS		
5.1	Conclusion	91
5.2	Future Works	94
REFERENCES		96
APPENDICES		
Appendix A	Questionnaire	100

LIST OF TABLES

Table No.		Page
Table 1.1	Research questions and research objective	6
Table 2.1	International data security standards	12
Table 2.2	Methods of sensitive data identification	18
Table 2.3	Comparison of sensitive data models	23
Table 2.4	Comparative analysis of model design concept	28
Table 2.5	Comparative analysis of the overall framework of the model	30
Table 2.6	Analysis of model limitations	33
Table 3.1	Expert information	51
Table 4.1	Model evaluation points	85

LIST OF ILLUSTRATIONS

Figure No.		Page
Figure 2.1	The definition of sensitive data (Ray, 2022)	15
Figure 3.1	The proposed comparison process of existing sensitive data models	41
Figure 3.2	Life cycle of data (Guo, 2024)	44
Figure 3.3	The proposed process of building a sensitive data governance model	44
Figure 3.4	Proposed sensitive data governance model based on the analysis of existing models	46
Figure 4.1	The proposed data security governance model	57
Figure 4.2	Area of expertise	79
Figure 4.3	Number of working years	79
Figure 4.4	Data asset comprehensiveness	80
Figure 4.5	Data management process	80
Figure 4.6	The novelty of data control technology	81
Figure 4.7	Data breach lifecycle coverage	81
Figure 4.8	Data breach lifecycle coverage	82
Figure 4.9	Data breach prevention effectiveness assessment	82
Figure 4.10	Comprehensive data security risk monitoring	83
Figure 4.11	Data asset risk coverage	83
Figure 4.12	Closed-loop management of data security risks	84
Figure 4.13	Land ability of the model	84

CHAPTER I

INTRODUCTION

1.1 RESEARCH BACKGROUND

As the global emphasis on data continues to increase, the field of data security, especially in the financial sector, is becoming increasingly valued for data security and privacy protection. In the financial sector, financial data primarily includes sensitive information such as personal identity information, transaction records, and financial information (Darem, 2023).

At present, all countries in the world have adopted corresponding information security measures to protect the data security of enterprises, and also attach great importance to the privacy protection of personal information. Enterprises hope to enhance their data security capabilities to help them better enhance the value of data and create more economic benefits. As a result, financial institutions around the world are increasingly focusing on data security, further driving innovation in the financial sector.

Financial institutions play an important role in protecting investor privacy and managing risk, so the security of customers' sensitive data is important (Lam, 2023). Sensitive data includes personal information as well as key information related to the investor's identity and finances, such as name, address, phone number, credit card number, and account balance. Once this customer data is leaked, it may be used by attackers for fraud and fund theft, which seriously threatens the security of investors' property (Chua, 2023).

As an information-intensive industry, the financial sector leverages a broad of

data to drive trading, risk management, and decision-making. From customer personal information and transaction records to market analysis, data plays a pivotal role in the operations of financial institutions. Consequently, the management and security of sensitive data are core tasks that financial institutions cannot afford to overlook.

While the use of data brings huge benefits to the financial industry, it also creates a great risk of data security leakage. With the continuous development of information security technology, network attacks, permission abuse and data leakage will make enterprises face a lot of data security risks. In addition, traditional cyber security protection is no longer sufficient to deal with new data security threats, so financial institutions need to establish stronger and more flexible data security control mechanisms to protect the data security of enterprises.

At present, financial enterprises are facing double pressure from internal threat and external threat. External threats include malware, viruses, ransomware and data breach attacks, all of which pose a huge threat to enterprise data security all the time. Insider threats, on the other hand, include employee access control, security awareness, and data breach issues, which put data at risk of active leakage.

To address the risks and challenges of data breaches, financial institutions still rely primarily on traditional cyber security measures such as firewalls, intrusion detection systems, and antivirus software. However, these traditional cyber security facilities have certain limitations in dealing with data breach threats and database attacks. For example, the firewall can only monitor and protect the attack behavior of network traffic, but it cannot detect and alarm sensitive data leakage. Therefore, with the continuous development of network attack technology, traditional security measures are not enough to deal with the risk of data leakage and attack behavior, enterprises need to establish protection measures and management mechanisms for data leakage risk.

1.2 PROBLEM STATEMENT

With the rapid development of information technology, data security risk management has become the focus of information security field. This section mainly discusses the problems and challenges of existing data security governance models, so as to provide guidance for the following research and practice of this study.

The risk of data breach is one of the major attack risks in the field of data security, which can not only cause financial losses to enterprises, but also damage the reputation and trust of enterprises and customers. Therefore, building an effective data security governance model is critical to protecting an enterprise's data assets. Although many data governance models have been proposed in previous studies, there are still many problems and challenges in the practical application of these models.

In this section, three major problems of existing data security governance models are analyzed in detail, which are the lack of model feature recognition and comparison methods, the lack of effective data security governance models, and the lack of effective model validation and evaluation mechanisms. These problems make the existing data security governance model have some limitations, but also increase the risk of enterprise in data security management. The purpose of this research is to build an effective data security governance model through the in-depth discussion of these problems, so as to reduce the risk of data leakage.

1.2.1 Deficiency of identification and comparison methods for limitations

There are many types of current data security governance models, but these models lack an effective method to identify their characteristics, which makes it difficult to effectively compare the models. Therefore, enterprises have great difficulties in selecting and implementing these data security models. For example, it is difficult for enterprises to determine which model best meets their specific data security needs.

On the other hand, as the amount of data managed by enterprises continues to grow and data types become more and more complex. Different enterprises may

encounter different data security risks, so enterprises need to choose the most appropriate data security governance model according to the specific situation (Ray, 2022).

However, there is a lack of unified model identification criteria in the literature, which makes it impossible for financial enterprises to judge the effectiveness of models in the process of evaluating and selecting models. Therefore, an effective feature recognition method needs to be established to help enterprises better understand and choose the appropriate data security governance model.

1.2.2 Lack of effective data security governance models for mitigating data breach risks

In the financial sector, data breaches not only lead to significant financial losses, but can also trigger a crisis of customer trust and legal liability (Abdullah, 2020). While many financial institutions have come up with data governance models, these tend to focus on the collection, storage and use of data at the expense of data breach prevention.

1.2.3 Absence of effective mechanisms for validating data security governance models

The effectiveness of data governance needs to be ensured through regular verification and evaluation (Asif, 2023). However, many financial institutions currently lack such mechanisms, making the effectiveness of data governance measures difficult to measure. At the same time, the effective implementation of data security will improve the company's brand reputation and enhance customers' trust in the company. Therefore, efficient data security control measures can effectively reduce the economic losses caused by data leaks and attacks, and ensure the sustainable development of the company.

1.3 RESEARCH QUESTIONS

As a data-intensive industry, the financial industry has a particularly prominent data security problem. Based on the problem, the research questions will delve into the following aspects:

1. What are the limitations of the existing data governance model?

First, the research focuses on the limitations of existing sensitive data governance models and conduct a comparative analysis of their effectiveness across various dimensions, including security, compliance, risk management capabilities, and data lifecycle protection. Through the evaluation of these models, which can identify the strengths and weaknesses of existing data governance frameworks.

2. How to build an effective sensitive data governance model to prevent data breach?

Secondly, the study focuses on how to propose a more effective data governance model through the shortcomings of existing data governance models. The main problem addressed is how to take appropriate measures to protect it from unauthorized access, data leakage and data misuse. This involves the entire life cycle of data, including its creation, storage, use, sharing, archiving, and destruction.

3. How to evaluate the proposed data security governance model in an effective way to be implemented?

In the third question, the research explores the regulatory measures used to protect data security in the financial industry, as well as the effects and potential limitations of these measures in practical applications. This includes understanding current security models such as multi-factor authentication, endpoint security, cybersecurity, blockchain technology, and more, and how to help financial institutions protect customer data and secure transactions.

1.4 RESEARCH OBJECTIVES

The table 1.1 presents three sets of research questions and research objectives, and focuses on how to propose a more effective sensitive data governance model to prevent financially sensitive data leakage.

Table 1.1 Research questions and research objectives

No.	Research Questions	Research Objectives
1	What are the limitations of the existing data governance model?	To identify the limitations of existing data governance models.
2	How to build an effective sensitive data governance model for the financial sector?	To design a sensitive data governance model that prevent sensitive data breaches in the financial field.
3	How to evaluate the proposed data security governance model in an effective way to be implemented?	To validate the proposed data governance model in the financial field by performing expert validation.

1. To identify the limitations of existing data governance models

The first objective is to identify the limitations of existing sensitive data governance models. By comparing and analyzing the effectiveness of different existing sensitive data governance models in various aspects, including security, compliance, risk management capabilities and data lifecycle protection capabilities. The advantages and disadvantages of existing data governance models are further found.

2. To design a sensitive data governance model that prevents sensitive data breach in the financial field.

The second objective is mainly to design a sensitive data governance model applicable to the financial field to effectively manage the sensitive data identified through the management of the various stages of the process.

3. To validate the proposed data governance model in the financial field by performing expert validation.

The third objective is to validate the proposed sensitive data governance model in terms of effectiveness, security, applicability. Through the method of expert survey to evaluate the technical architecture and practicality of the proposed model.

1.5 RESEARCH METHODOLOGY

This study employs the following research methods to conduct the research work:

1. Case Study

Through in-depth analysis of existing sensitive data governance case studies, this study explores the characteristics and effectiveness of data governance models from multiple angles, revealing the performance and limitations of different models in actual applications.

2. Model Comparison

Based on the identification of shortcomings in existing models, this study has developed a more effective data governance model. The model aims to comprehensively protect data from data breach risks throughout its entire lifecycle.

3. Questionnaire survey

By designing expert questionnaires, this study collected data on the existing model risk mitigation measures from security experts at different financial institutions. Through statistical analysis of these data, it provided important reference for model improvement.

1.6 RESEARCH SCOPE

The scope of this study is confined to data breach risk prevention in the financial sector, with a specific focus on the potential data breach risks encountered during the data lifecycle, including collection, identification, processing, and exchange. By conducting a comparative analysis of existing sensitive data governance models, this study effectively identifies their limitations and characteristics. Building on this foundation, the study develops a new data security governance model designed to effectively manage and protect sensitive data in the financial sector.

Additionally, the study validates the effectiveness, security, and applicability of the proposed governance model through expert data security risk assessments. Through this comprehensive process, this study aims to provide the financial sector with a more robust data security solution.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

The literature review section is analyzed by searching the literature resource website for articles in the following areas, including: criteria for sensitive data, detection of sensitive data, and sensitive data governance models. The research employs a targeted search methodology, utilizing keywords such as 'Financial Data Security', 'Sensitive Data Identification', 'Data Governance Models', and 'Data Privacy in Finance', to systematically explore the existing body of literature.

The research first discusses the definition of sensitive data in the financial field and the technical means of identifying sensitive data proposed in the literature, such as natural language processing and machine learning, to identify and classify sensitive data in the financial industry. The dimensions of analysis include literature background, research questions, research objectives and research methods. The effectiveness and potential application prospects of sensitive data identification technology are emphasized (Wang, 2022).

Regarding data governance frameworks in the financial sector, existing literature primarily centers on the conceptual underpinnings and practical applications of data governance, encompassing elements such as data acquisition, storage, utilization, and dissemination. The analytical dimensions include the contextual backdrop of the literature, specifically the necessity for data governance within the financial industry; the research inquiries, which concentrate on the obstacles encountered in the implementation of data governance in the financial domain; the

research objectives, which aim to establish a robust governance framework to guarantee data security and regulatory compliance; the research methodologies, which incorporate qualitative analysis and case studies; and the principal findings, which highlight the advantages and drawbacks of various governance models (Chua, 2023).

About the validation of data governance models, the study analysis the application and effectiveness of different models in the financial industry. The dimensions of analysis include the background of the literature, that is, the need to compare different data governance models; research questions, exploring the applicability and efficiency of different models; research objectives, comparing and evaluating the performance of different models; research methods, using comparative analysis and evaluation indicators; and main results, providing comprehensive evaluations of different models (Yebenes, 2021).

In summary, the literature review of this study synthesizes key research in the field of sensitive data governance in finance, revealing the latest advancements and challenges in automated identification technology, governance framework construction, and comparative analysis of models, providing theoretical support and guidance for data governance practices in the financial industry.

2.2 THE CLASSIFICATION OF SENSITIVE DATA IN THE FINANCIAL SECTOR

2.2.1 Data security standards

As early as 1995, the European Union developed and enacted the Data Protection and Privacy Law, which set out the definition of personal data, the methods of processing, storage and transmission, and the responsibilities and obligations of data protectors and data processors (Tzanou, 2020). With the popularity of the Internet and the rise of mobile payment, data security has also become an important issue in the field of regulatory system, and countries around the world have begun to formulate relevant laws and regulations to protect the data security of individuals and organizations.

California made a law called the California Consumer Privacy Act to keep people's private information and consumer rights safe. The law tells how people in California can control their own information, like how it's collected, used, shared, and kept safe (Stallings, 2020). This means people have more control over their information, and businesses have to follow higher standards when they deal with personal information.

Hale ML (2022) pointed out that NIST made a rule called NIST SP 800-53. This rule has many suggestions to help organizations make their information security better and work more efficiently. It also helps in making and putting into action information security rules and plans.

The group in charge of payment cards, called the Payment Card Industry Security Standards Council, updated a rule called PCI DSS 4.1. This new rule focuses on keeping sensitive data safe, like credit card information and personal details (Kozminykh, 2022). The rule talks about things like how to encrypt data, control who can access it, and how to get data back if needed, to make sure payment information is secure.

Tzanou Maria (2020) wrote in the literature that the European Union and its member countries made a law to protect personal data from being misused, lost, or revealed. This law, called the Data Protection and Privacy Act, explains what personal data is, how it can be processed, stored, and sent, and what the people in charge of protecting and handling data need to do.

The fact that the EU had further strengthened data protection regulations mentioned by Akshar (2021). The General Data Protection Regulation (GDPR) had been enacted. The GDPR was designed to protect all personal data from unauthorized processing and exploitation, covering customer information recorded electronically or otherwise, confidential information, transaction information, location data and device data, which can be used to identify and contact customers and process their personal information.

The European Union passed the Cyber security Act (Khurshid, 2022). The first EU-wide cyber security certification scheme has been established. This is of great significance for the construction of network and information and communication security systems in EU member States, as well as the improvement of security risk prevention and control capabilities.

The EU and its member States published the European Data Strategy and the Data Governance Act (Botrugno, 2023). The European Data Strategy aimed to achieve the vision of a true single data market, addressing through policy and financial measures the problems identified on the basis of previous achievements. The Data Governance Act strengthened the European mandate on public data and lay the foundation for a new approach to data governance in Europe.

The International Organization for Standardization (ISO) and the International Telecommunication Union (ITU) jointly released the ISO/IEC 27001 standard, which was mentioned by Carovano (2023). The standard covered all aspects of information security management, including security policies, security programs, security incident management, security assessments, and security controls to protect an organization's information assets from unauthorized access, use, modification, and disclosure.

Table 2.1 gives the international laws and standards issued in the field of data security in the past 20 years, and introduces the organization and time of promulgation of relevant standards.

Table 2.1 International data security standards

No.	Issuing Organization	Laws and Regulations	Year
1	European Union and European Countries (Botrugno, 2023)	European Data Strategy	2020
2	European Union and European Countries (Botrugno, 2023)	Data Governance Act	2020
3	European Union and European Countries (Khurshid, 2022)	Cybersecurity Act	2019
4	State of California, USA (Stallings, 2020)	California Consumer Privacy Act	2018

to be continued ...

...continuation

5	Payment Card Industry Security Standards Council (S Kozminykh, 2022)	PCI DSS 4.1 Industry Standard	2018
6	National Institute of Standards and Technology (NIST) (R Kudo,2022)	NIST SP 800-53	2017
7	European Union and European Countries (Akshar, 2021)	General Data Protection Regulation (GDPR)	2016
8	International Organization for Standardization (ISO), International Telecommunication Union (ITU) (Carovano, 2023)	ISO/IEC 27001	2013
9	European Union and European Countries (Tzanou, 2020)	Data Protection and Privacy Law	1995

By summarizing major data security laws and standards, this study presents a timeline for the release of data security regulations, providing an effective understanding of different data security compliance requirements. At the same time, this requirement serves as the guidance of data security compliance and the requirement of compliance part, which helps this research to better build the data security governance model.

2.2.2 Definition of sensitive data

In recent years, there have been many references to the definition of sensitivity. Chua (2023) highlighted the classification of personal data and its impact on privacy concerns and propensity to disclose information. By investigating and analysing the characteristics of personal information practices, he studied the differences in privacy concerns and information disclosure intentions in different categories, thus identifying different categories of personal data.

Darem (2023) emphasized the importance of the banking and financial sector as a primary pathway for cyber attacks and proposed the goal of classifying cyber threats and describing effective countermeasures. They used a categorical approach to the severity and technical complexity of cyber threats.

The necessity of protecting the confidentiality and integrity of sensitive information in cloud systems was emphasized by Alotaibi (2022), which focused on the description of sensitive data categories and the definition of attack scope. It illustrated techniques to mitigate the risks of data storage in cloud infrastructure, and finally discussed the dilemmas and challenges of sensitive data exposure.

Aljumah and Ahanger (2020) mentioned the need for strict security protocols as cloud computing services become more dependent. Their goal was to dissect security challenges in cloud computing environments and propose strategies to mitigate these threats. The research methodology included a comprehensive survey of the inherent vulnerabilities and countermeasures of cloud computing.

The common point of the above literature was that they all focused on the problem of data security and privacy protection, and adopt classification and analysis methods to study the problem, and all aimed to propose solutions or countermeasures for sensitive data definition. They differed in the areas and typed of data they focused on. For example, Chua (2023) focused on the classification of personal information and privacy concerns, Darem (2023) focused on cyber threats in the banking and finance sector, Alotaibi (2022) focused on the protection of sensitive data in the cloud environment, and Darem (2023) extensively discussed the security challenges of cloud computing. In addition, the research methodology and proposed solutions for each article differ, reflecting the specific needs and challenges of their respective research fields.

Illustrated in Figure 2.1, the definition of sensitive data can be divided into different aspects.

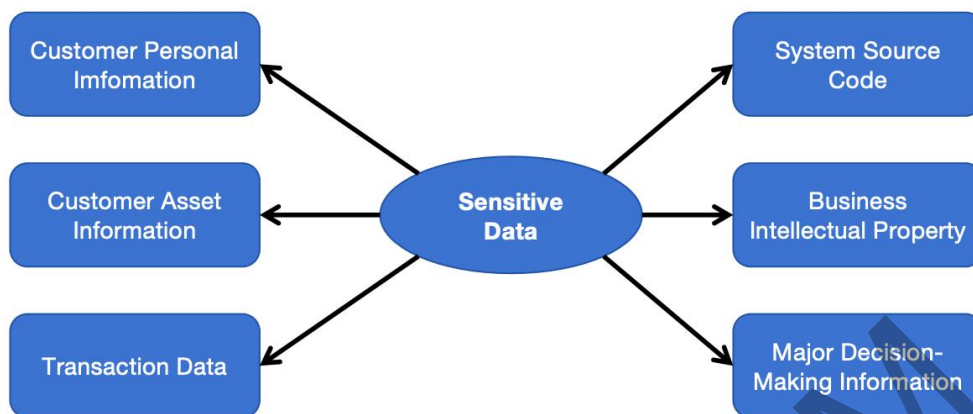


Figure 2.1 The definition of sensitive data (Ray, 2022)

1. Customer Personal Information:

This includes all personal information of the customer, such as name, contact information, account details and transaction records. These pieces of information are directly related to the customer, and if leaked, could lead to serious information security and privacy leakage issues.

2. Customer Asset Information:

Generally refers to all asset information of the customer, such as account fund status, position details and transaction details, directly reflecting the customer's financial and asset information, which requires strict protection.

3. Transaction Data:

The company's real-time transaction data and historical databases reflect market liquidity and market conditions, have significant commercial value, and unauthorized disclosure can undermine fair competition.

4. System Source Code:

The source code of the core trading systems and risk control systems of financial companies is particularly important. The source code is an intellectual property asset with extremely high commercial value and is the foundation for maintaining the company's core competitiveness.

5. Business Intellectual Property:

Business analysis models, forecasting models, and other intellectual properties generated through means such as big data analysis, once leaked, can cause irreparable strategic losses.

6. Major Decision-Making Information:

Important decision-making information of the management team, such as executive meeting minutes, restructuring plans and planning reports. The leakage of this information can severely affect the company's reputation and confidence in the capital market.

According to the classification results of the existing sensitive data, the above data is divided into three levels, which are high sensitive, medium sensitive and low sensitive. Among them, highly sensitive data refers to information related to customers, including customer personal information and customer asset information, and sensitive data includes transaction data, system source code and Business Intellectual Property. Low-sensitive data is Major Decision-Making Information.

2.2.3 Methods of sensitive data identification

In the digital age, the identification and protection of sensitive data has become an important issue. The following reviews several literature on sensitive data identification methods and tools, showing the contributions and achievements of different researchers in this field.

A new concept, Sensitive Privacy (SP), which aimed to solve the problem of outlier analysis while protecting privacy (Asif, 2023). Because the existing definition of Privacy did not guarantee accuracy when performing outlier analysis, SP provided strong privacy protection like Differential Privacy (DP) while being able to analyse data with actual meaningful accuracy.

The study has developed a novel n-step foresight mechanism to efficiently answer arbitrary outlier queries and demonstrated that this mechanism can guarantee sensitive privacy when restricted to a common class of outlier models. In addition, the researchers provided general methods for building sensitive privacy mechanisms and show under what conditions these constructs are optimal.

In the financial cloud environment, Zu (2024) proposed a sensitive data classification method called UP-SDCG (Unionpay-sensitive data classification and grading) that specifically targets collaborative edge computing. This method realized the automatic classification and grading of financial sensitive structured data by constructing hierarchical classification database and using database enhancement technology and synonym identification model.

Experimental analyzed on simulated datasets show that UP-SDCG achieved an accuracy of more than 95%, which was better than the other three comparison models. In addition, the method was tested in real-world financial institutions and achieved good results in customer data, regulatory and personally sensitive information.

Peng (2022) proposed a way to find sensitive data using deep learning because edge technologies are becoming more popular and we need to protect sensitive data. They first used a Bert model and patterns to get sensitive data from different types of data. Then, they suggested a framework that uses both the Bert model and patterns to find sensitive data with high accuracy and good generalization. The experiments showed that the plan worked well.

A data leak prevention (DLP) system based on file fingerprint similarity detection was proposed by Wang (2020). The system used a method based on the

SimHash algorithm to find similarities between document fingerprints. They presented three types of SimHash algorithms for getting features: keyword-based (KbS), paragraph-based (PbS), and a special paragraph-based (SoP) fingerprint. They calculated the distance between fingerprints by making file fingerprints and comparing them to a library of sensitive documents.

Nikoletos (2023) proposed a method to automatically find and hide sensitive data, called RoG. This method uses natural language processing (NLP) technology and can work with many data sets and different areas of sensitive data governance. RoG uses NLP techniques like named entity recognition (NER) and patterns to find sensitive data, and then replaces it with fake names or general terms using techniques like the K-anonymous algorithm.

A new context-based method for identifying sensitive data, known as CASSED (Context-based Structured Sensitive Data Detection Method), had been proposed by Kužina (2023). This method uses BERT technology based on transformers and natural language processing (NLP) to find sensitive data. Compared to old methods that use rules, CASSED finds sensitive data more accurately by looking at the relationships between cells in the same column and the context between different columns.

As shown in Table 2.2, it presents existing methods for sensitive data identification, and list the goals achieved by different methods and their main contributions.

Table 2.2 Methods of sensitive data identification

Methods	Title	Author	Year	Objective	Contribution
UP-SDCG	UP-SDCG: A Method of Sensitive Data Classification for Collaborative Edge Computing in Financial Cloud Environment	Lijun Zu	2024	Strengthening financial data security and compliance with regulatory requirements	Enhancing the precision and efficiency of data classification

to be continued ...

...continuation

A method of identifying anomalies	Identifying Anomalies While Preserving Privacy	Hafiz Asif	2023	Achieving a balance between data privacy protection and anomaly detection	Accurately analyzing anomalies in data while maintaining privacy
RoG§	RoG§: A Pipeline for Automated Sensitive Data Identification and Anonymisation	Sotiris Nikoletos	2023	Protecting individual privacy and preventing data breaches	Identifying and anonymizing sensitive data in large datasets
CASSED	CASSED: Context-based Approach for Structured Sensitive Data Detection	Vjeko Kužina	2023	Enhancing privacy protection and automating the sensitive data identification process	Detecting sensitive data in structured data
A method of Deep Learning	Deep Learning Based Sensitive Data Detection	Peng Chong	2022	Achieving a sensitive data detection method with high accuracy and generalization capability	Real-time identification of sensitive data from structured and unstructured data
A method of file fingerprint identification and detection	Application Research of File Fingerprint Identification Detection Based on a Network Security Protection System	Chunwei Wang	2020	Enhancing enterprise information security and preventing data leakage	Preventing the leakage of corporate confidential documents

According to the comparison of existing sensitive data identification methods, this study shows the research focus of existing sensitive data identification methods, so as to help the sorting of existing sensitive data assets and classified management of assets.

2.3 COMPARISON OF EXISTING SENSITIVE DATA GOVERNANCE MODELS

In the area of sensitive data governance, there has been a lot of literature in recent years proposing different kinds of approaches to address the challenges and opportunities presented by the era of big data.

Yebeles (2021) proposed a sensitive data governance framework for Industry 4.0, which focused on three core aspects: Data as a Service (DaaS), Platform as a Service (PaaS), and Monitoring as a Service (MaaS). The model had steps like finding out what was needed, putting data together, making rules for handling data, watching how things were done, and making things better over time. The big issue was how to manage sensitive data well in the Industry 4.0 environment to make sure data assets were used and managed effectively.

A sensitive data governance model was proposed by Lee (2020). The steps included managing data, planning data structure, combining data, setting up a smart data warehouse, and managing data quality. This plan solved how to make supply chain processes more efficient and reliable by combining data and using business analysis. It proved the model worked well by using it in a textile company in Peru.

The fact that a conceptual framework for data governance was developed through a systematic review of the literature (Wirtz, 2022). The study looked at 145 studies and reports from 2001 to 2019. The study aimed to give a full view to help people working on data governance. The goal was to find the main parts of data governance and break them down into six parts for future research.

Wang (2022) designed and implemented a big data platform for a large hospital at West China Hospital of Sichuan University. The steps included getting data from different places, putting it together, managing it, checking the quality of data, and making a data security system. The study solved how to make different kinds of data from many places work together and be safe.

The platform used a master-slave model to bring together a lot of different data quickly and set up a place to keep different kinds of data separate for storage and computing. The goal was to make a big database with lots of data to help with patient care, research, and management.

An ontology-based data governance model, aimed at simplifying the complexity of big data environments, has been proposed by Castro (2021). The steps

included making an ontology, developing a system with parts that can work on their own, sharing knowledge, and using automatic reasoning with ontologies. The study looked at how decisions about data governance could be controlled with meaning and rules.

The study used reasoning with ontologies and a system that could work on its own to make data governance smarter and more automatic. The goal was to make managing big data simpler and to show the technology worked by using a test system in a global video service.

Benson (2023) proposed a host-based Data Breach prevention (DLP) approach that protects data by tracking the flow of sensitive information in a file system. The idea was that any sensitive data going to a non-sensitive place would make that place more secure. By marking places with sensitive data, this method made sure sensitive information wasn't stored in the wrong places, stopping information leaks through hiding, changing, compressing, or encrypting data. This method was put into the Linux operating system as a special part and could work with older programs.

The study by Guo (2024) focused on recovery strategies after data breaches, particularly the relative effectiveness of functional and financial remedies. Through two scenario-based experiments, the researchers found that functional remedies were more effective than financial remedies when sensitive information was compromised. Functional remediation directly affected the client's negative coping behaviours and indirectly affects those behaviours by reducing feelings of fear and anger. Financial remediation affected negative behaviours indirectly by reducing anger, but has no direct effect on reducing fear. The study provided key insights into how to manage customer responses and recommends the use of well-designed data breach recovery strategies.

Du and Shu (2023) designed and implemented a deep learning-based financial risk monitoring and early warning system for China. The system aimed to provide financial regulators with a sensitive and scientific financial risk early warning system under the COVID-19 pandemic. Through the deep learning model, the system could

detect the financial risk behaviours brought by a few people, which provided a new theory and method for the financial risk management system.

A framework for predicting the risk of sensitive data breaches was proposed by Fang (2021). This idea used how different time series depend on each other to deal with not having enough data. The main idea was to use a special structure to see how time series depend on each other and to find the model's settings with a two-step method. When this idea was used with data from companies, it helped to find out the risk of sensitive data being leaked. Tests showed that this idea was good at guessing the risk of data leaks.

Jan (2020) proposed a systematic asset classification model for security testing. The model looked at how open systems were, how important the data was, and how much security was needed. Researchers looked at 451 information systems and picked them for safety tests to see if data could be leaked. The model was good at giving different levels of security to things in the test and finding possible safety holes.

The sensitive data identification technology of industrial Internet based on traffic analysis was studied by Bi (2020). The study suggested a new model that combined the special features of sensitive data in the industrial Internet with traffic analysis. This helped to find sensitive data in the industrial Internet well. The model had steps like collecting data streams, keeping them, analyzing them, and giving results. Also, by looking deep into message content and matching them with rules, the model could find and judge sensitive data flows well.

Table 2.3 presents a comparison of the different existing sensitive data models, including their advantages, disadvantages, contribution, and major objectives.

Table 2.3 Comparison of sensitive data models

Model Name	Authors	Advantages	Disadvantages	Contribution	Objectives
Recovery Strategies after Data Breach	Yuanyuan Guo, Chaoyou Wang, Xiaoting Chen	Differentiates effectiveness of functional vs. financial remedies, considers customer emotions like anger and fear	Limited to data breach recovery scenarios, does not cover preventive measures	Managing customer reactions after a data breach	Suggesting appropriate recovery strategies to mitigate negative customer behaviors post-breach
Enhancing Zero Trust Models through Blockchain Integration	Clement Daah, Amna Qureshi, Irfan Awan, Savas Konur	Integrates blockchain for enhanced security, addresses identity, access, data protection, and network security	Potential scalability and adaptability challenges with blockchain integration	Cybersecurity in the financial industry, particularly against APT attacks	A comprehensive security architecture that mitigates cyber threats and builds consumer trust
China Financial Risk Monitoring and Early Warning System	Peng Du, Hong Shu	Based on deep learning, detects financial risk behaviors, provides new theory and method for FRM	High data requirements for training models	Financial risk management, especially in the context of economic changes like COVID-19	Establishing a sensitive and scientific financial risk pre-alarm system
Host-based DLP Method	A. L. Lamidi Benson	Tracks sensitive data flow within file system, prevents covert channels for information leakage, works with legacy applications	May produce false positives, potential tag flooding over time	Data leak prevention, preventing sensitive data from residing in non-sensitive objects	Prevent data leakage through various techniques such as steganography, data modification, compress, or encryption
Big Data Health Care Platform	Miye Wang	Integrates multisource heterogeneous data, ensures data security, supports large-scale data processing.	High initial investment in technology, complexity in managing diverse data types.	Challenges in integrating, calculating, storing, and governing multisource heterogeneous data in healthcare.	To build a high-quality data asset platform that supports clinical activities, scientific research, and management.

to be continued ...