北京金融科技产业联盟
BEIJING FINTECH INDUSTRY ALLIANCE

金融业智能数据脱敏技术研究报告

北京金融科技产业联盟

2025 年 4 月

版权声明

本报告版权属于北京金融科技产业联盟，并受法律保护。转载、编摘或利用其他方式使用本报告文字或观点的，应注明来源。违反上述声明者，将被追究相关法律责任。

编制委员会

编委会成员：

何　军　　聂丽琴　　童　蕙

编写组成员：

曹嘉欣　　王仰东　　周晓阳　　陈永康　　单姜一　　白　梅

焦　航　　王志远　　赵天蔚　　韩韶欣　　谭贵强　　闫　瑾

温国梁　　韩明宵　　丰　瑾　　张晓玉　　许江峰　　林　宇

彭　晋　　白晓媛　　杜晓黎　　姜志辉　　杜啸争　　李　娜

张邦军　　乔文汇

编　审：

黄本涛　　国　钰　　魏中宣

参编单位：

交通银行股份有限公司

华为技术有限公司

中国邮政储蓄银行股份有限公司

蚂蚁科技集团股份有限公司

中电金信软件有限公司

目 录

摘要：金融业在数字化转型进程中面临着数据安全和隐私保护的严峻挑战，亟需对敏感数据进行精细化管控，实现实时性、多样性的数据脱敏处理。本课题围绕金融业智能数据脱敏技术的发展现状和行业实践，深入探索并剖析数据脱敏技术在金融行业的具体要求，归纳总结智能数据脱敏的关键技术，推进智能数据脱敏在金融行业数据安全领域的深入研究和场景应用，为金融机构在数字化转型道路上筑牢数据安全防线提供参考与支撑。

一、总体概述

（一）研究背景

在21世纪的信息化浪潮中，金融行业与大数据的结合日益紧密，数字化转型成为推动金融创新的核心动力。金融行业积累了海量用户数据，这些数据不仅包括金融产品信息、客户服务记录，还涵盖了个人身份信息、消费习惯等敏感数据。这些数据的积累，一方面为金融行业提供了精准营销、风险控制、客户服务等场景的决策支持，另一方面也带来了严峻的数据安全和隐私保护挑战。

金融行业较其他行业其特殊性在于数据具有更高的敏感性。个人账户信息、交易记录等数据一旦泄露，不仅会侵犯到用户隐私，还可能导致金融诈骗、资产损失等严重后果。因此，在利用数据推动业务发展的同时，确保数据安全和用户隐私，成为金融行业亟需解决的问题。

随着全球范围内对数据保护意识的增强，各国政府纷纷出台了相关的法律法规，如欧盟的通用数据保护条例（GDPR[1]）、美国的加州消费者隐私法案（CCPA[2]）等，中国也相继颁布了《中华人民共和国网络安全法》《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》，对个人信息的收集、处理和使用提出了严格的要求。这些法规的实施，对金融机构的数据管理

[1] GDPR，全称《通用数据保护条例》(General Data Protection Regulation)，由欧盟委员会创立，是目前为止在欧洲乃至全球最严格的个人数据隐私保护法规之一。

[2] CCPA，全称《加利福尼亚消费者隐私法》(California Consumer Privacy Act)，是美国加利福尼亚州的一部数据隐私法，适用于所有收集、处理或出售加州消费者个人数据的企业。

提出了更高的标准，金融机构必须采取有效的技术手段，确保在遵守法律法规的同时，保护数据的安全和隐私。

在金融行业内部，数据的共享和交换是常态，数据在多个部门间流动，如安全部门、测试部门、业务部门、数据部门等。内部的数据流动虽然有助于提高工作效率，但也增加了数据泄露的风险。因此，金融行业需要在数据的传输、存储和共享过程中，采取有效的数据脱敏措施，以降低数据泄露的风险。金融业的数据脱敏需求也体现在开发与测试场景中。在开发测试过程中，为了高度模拟生产环境，需要导入大量用户真实数据分析处理。为保护用户敏感信息，如姓名、身份证号等，避免在此过程中发生数据泄漏，需要对生产环境中的用户敏感数据进行替换、匿名等脱敏操作。

综上所述，在遵守法律法规的同时，有效保护数据安全和用户隐私，是金融行业亟需研究和解决的问题。数据脱敏技术作为解决这一问题的有效手段，其研究和应用成为了金融行业数据安全管理的重要组成部分。

（二）数据脱敏的重要性

在数字化时代，金融行业将数据视为核心资产，并不断探索其价值与潜力，但大数据的广泛应用也带来了数据安全和隐私保护的挑战。数据脱敏技术之所以重要，是因为它在保护个人隐私和企业数据安全的同时，确保了数据的可用性和业务的连续性。

1.遵守合规要求

数据脱敏是遵守法律法规的基本要求。随着全球数据保护法规的出台和完善，企业必须采取有效措施来保护个人数据。违反这些规定可能导致重大的法律风险和经济损失。数据脱敏技术能够帮助企业在处理个人数据时，避免直接处理敏感信息，从而降低违法风险。例如，通过掩码处理身份证号码和银行账号等敏感信息，企业可以确保在遵守法律法规的同时，保护客户数据不被泄露。

2.维护企业权益

数据泄露事件不仅损害客户利益，还严重影响企业的公众形象和市场信誉。通过实施数据脱敏，企业能够在内部管理和外部交互中保护客户数据，增强客户对企业的信任，从而在激烈的市场竞争中保持优势。例如，企业在进行客户服务时，使用脱敏数据可以确保即使数据被泄露，也不会暴露客户的敏感信息，从而保护客户权益和企业声誉。

3.保护数据安全

金融行业数据量大，具有高价值，是网络攻击的主要目标。数据脱敏能够降低敏感数据在存储、处理和传输过程中的风险，减少数据泄露的可能性。这对于防范内部威胁和外部攻击，保护企业免受数据泄露的影响至关重要。例如，在金融行业的运维管理中，系统运维人员由于具有较高权限成为数据泄露的主要风险源，因此需要对核心敏感数据进行脱敏，预防数据泄露。

4.提升数据可用性

脱敏后的数据可用于数据分析、机器学习等业务场景，支持企业决策和业务创新。这使得企业能够在保护隐私的同时，充分利用数据资源，实现数据价值的最大化。例如，在金融风控场景中，通过对支付信息进行脱敏处理，可以在不泄露用户敏感信息的前提下，识别用户账户是否存在盗冒用的操作风险，提升风控的精确性和有效性。

5.支撑数字化转型

随着金融服务日益依赖数据驱动的解决方案，数据脱敏技术成为保障数据安全与隐私的关键支撑，为金融行业的可持续发展奠定了坚实基础。以金融授信场景为例，通过对账户信息、转账记录及交易支付数据进行数据脱敏降级处理，不仅能够精准分析用户的还款能力，还能有效保护用户的敏感信息免遭泄露，在提升业务效率的同时，充分满足合规与隐私保护的双重要求。

6.推动技术应用

金融业的数据脱敏需求贯穿于开发测试、运维管理等核心环节，旨在对敏感数据进行全方位保护。面对数据安全和隐私保护的严峻挑战，数据脱敏技术作为一项关键解决方案，已成为金融行业数据安全管理体系的重要组成部分。例如，在金融营销场景中，通过对账户信息、交易支付信息等敏感数据进行脱敏，金融机构能够在保护用户隐私的前提下，精准评估用户的资产能力及风险意识，从而为用户推荐适配的金融产品与服务。这种脱敏技术的应用，不仅提升了数据使用的合规性与安全性，还在业务创

新与隐私保护之间实现了有效平衡，为金融业的可持续发展提供了坚实保障。

综上所述，数据脱敏技术在金融行业中不仅是应对合规挑战的必要手段，更是保护企业资产、维护客户关系、释放数据价值的关键措施。随着数据安全形势的日益严峻以及隐私监管要求的不断强化，数据脱敏的重要性将进一步凸显，成为金融行业不可或缺的技术保障。通过持续优化脱敏技术，金融机构能够在确保数据安全的同时，充分挖掘数据潜力，为业务创新与客户服务提供强有力的支持，推动行业在合规与效率中实现高质量发展。

二、应用现状

（一）应用场景

数据脱敏应用场景分为技术场景和业务场景，技术场景主要包括开发测试、数据分析、数据交换、数据共享、生产应用、运维应用等，业务场景包括营销获客、风险防控、业务经营等。

1.技术场景

（1）开发测试

在金融业务系统开发与功能升级中，涉及客户基本信息、社会关系、财产信息等敏感数据需被严格保密。开发测试阶段，必须确保测试数据在保持真实数据特征的同时，不泄露任何真实客户信息，这就需要运用脱敏技术来处理这些数据。

（2）数据分析

数据分析目的在于深入挖掘数据价值。在此过程中，关键的用户特征如年龄、性别、地区和行为记录等信息需被保留，而用户的敏感身份信息和非必要的敏感字段则需被脱敏，以确保研究数据准确性、有效性和安全性。

（3）数据交换

数据交换场景主要通过 API 接口方式，向金融业机构内部特定平台提供数据，数据请求时会附带用户信息，需要对部分用户信息进行脱敏。

（4）数据共享

面对特定的业务合作和联合营销需求，通过数据流转提升数据价值，数据共享变得不可避免。在这一过程中，根据业务目标的不同，对敏感数据采取相应的脱敏技术，如数据抑制和扰乱等，以确保数据在共享过程中的安全。

（5）生产应用

在生产系统中使用数据时，应根据业务需求确定用户对敏感信息的最小访问权限。在必须访问敏感信息的情况下，应通过掩码屏蔽等脱敏手段保护数据，以降低数据泄露风险。

（6）运维应用

运维人员虽拥有信息系统的高权限账号，但其主要职责是对数据库进行监控和审计，而非深入了解系统内部的具体数据。

2. 业务场景

（1）营销获客

在大数据时代，银行通过深度挖掘数据价值来增强竞争力，促进消费增长和用户活跃度。在利用第三方流量和触客优势营销时，需整合金融业务数据与外部数据，构建客户画像，实现精准营销和高效转化，提升客户消费体验。此过程中，保护客户个人信息至关重要，必须采用智能脱敏技术以防止敏感信息泄露。

（2）风险防控

金融行业建立跨机构风险信息共享机制，如银行卡风险信息共享和涉电信网络诈骗风险信息共享，以及信贷业务和征信报告

评分模型。这些措施基于内外部交易和历史数据，结合智能规则引擎，实时预测和分析欺诈等非法行为。在模型建设中，对敏感数据进行脱敏，确保数据“可用不可见”。

（3）业务经营

金融业务经营中，如交易账单处理、信用卡权益合作和快捷支付等场景，除向个人或组织展示数据或履行法定职责外，应优先对敏感数据进行脱敏处理，以保护客户隐私。

（二）建设进展

当前，金融行业在数据脱敏技术应用上取得显著进展。多数机构已建立起适应多场景需求的数据脱敏体系，通过标准化平台整合现有数据基础设施，并逐步覆盖开发测试、生产运维、数据共享等核心场景。

其中，静态脱敏与动态脱敏技术作为保护数据隐私的关键措施，已被广泛采用并趋于成熟。静态脱敏通过ETL（Extract-Transform-Load）工具批量处理非实时的数据，能够在保护隐私的同时，为数据分析和挖掘保留关键信息。这种方式常被业内用于搭建测试环境中，通过静态脱敏技术，在数据被导入测试环境之前，对其中敏感信息进行替换、加密等处理，能达到保留业务特征又符合隐私保护要求的效果。

动态数据脱敏则是通过中间件或数据库内核内置规则等方式实现。通过在生产系统对外查询时部署动态数据脱敏，能够在保护数据隐私的同时，确保数据请求的实时性和效率，有效降低

运维过程中高权限账户泄露的风险。完善的实时数据脱敏和监控系统，还能及时发现潜在的数据泄露风险，确保数据在其整个生命周期中得到全面的保护。

行业层面对于数据脱敏策略的标准化管理已呈现两个重要趋势：一是建立可编排的规则引擎，通过多层次策略体系（如机构级基础规则、系统级定制规则、接口级动态规则）实现字段级精准控制。二是自动化脱敏平台深度并集成多模态数据接口，通过自适应解析技术打通 Oracle、MySQL 等关系型数据库与 Hive、HBase 等大数据组件间的技术鸿沟，并通过 CSV、Excel 等结构化文件的智能解析，覆盖主流文件格式的批量脱敏需求。

在跨机构协作场景中，通过整合多方安全计算与联邦学习框架，各参与方能够在不暴露各自原始数据的情况下，共同完成数据分析任务，既满足了业务需求，又保障了数据的安全性。

对于金融行业来说，数据脱敏不仅要考虑如何有效地隐藏敏感信息，还需要保证脱敏后的数据依然保有其原有价值，以便后续分析和决策使用，其安全性、准确性和实用性成为了衡量其优劣的关键标准。

三、安全要求

本章从金融业保护敏感数据的重要性出发，基于国家及业内关于数据安全保护的规章制度，从数据服务、数据流通、数据管理三个角度，探讨了金融业对智能数据脱敏的要求，以确保敏感数据在使用、传递、留存时的安全性。

（一）数据服务安全要求

在数据服务层面，金融机构通过数据脱敏技术，构建了多层次的用户隐私与数据安全保障体系。在提供数据服务的过程中，数据脱敏能够显著降低数据泄露风险，确保敏感信息的安全性。数据服务主要涵盖业务数据查询、客户数据分析、投资风险管理等场景。

1.合规合法性

从“个人金融信息”[3]的概念提出，到金融业数据工作的五大基本原则[4]的确立，及至今年《银行保险机构数据安全管理办法》的征集[5]，不断完善和严格的法律规章制度体现出监管机构对金融敏感数据的愈发重视。这要求金融机构在使用数据提供服务前，如客户的身份信息、财务数据等敏感数据需要根据《中华人民共和国个人信息保护法》《中华人民共和国数据安全法》等法规在数据的原有形态上进行脱敏，以确保所见数据的合法、正

[3] 《关于银行业金融机构做好个人金融信息保护工作的通知》，由中国人民银行于 2011 年发布，是我国最早的官方对金融业敏感数据提出要求的文件。其中首次使用了“个人金融信息”这一概念，其法律依据是《中国人民银行法》《商业银行法》《反洗钱法》《个人存款账户实名制规定》等法律法规。

[4] 《金融业数据能力建设指引》，由中国人民银行于 2021 年发布，提出金融业数据工作的五大基本原则，包括用户授权、安全合规、分类施策、最小够用、可用不可见。

[5] 《银行保险机构数据安全管理办法》，由国家金融监督管理总局于 2024 年征集，旨在规范银行业保险业数据处理活动，保障数据安全和金融安全。

当。例如对身份证号码、银行账号等数据进行掩码处理，以降低泄露风险。

2.精细化数据控制

客户金融信息在完成数据分析后，可被应用于个性化推荐或金融数字产品的投资风险管理中。在此过程中，智能数据脱敏技术需严格遵循“最小必要原则”，实施精细化脱敏处理，仅提供完成业务目标所需的基本信息，并对非必要展示的敏感数据进行脱敏。例如，在对可疑交易进行反洗钱监控和审查时，系统传递给审查人员的交易数据需对客户的敏感信息（如住址、身份证号等）进行脱敏处理，确保仅必要的账户交易信息可被查看，从而在满足业务需求的同时，最大程度保护用户隐私。

此外，脱敏后的数据需保持统计代表性和分析有效性。这对智能数据脱敏技术提出了更高的要求：在确保数据业务特点得以保留的同时，还需对隐私信息进行有效保护。

3.实时性和多样性需求

在金融科技领域，快速和简便的操作是用户选择金融产品的标准，而“低用户感知”是确保用户体验和满意度的关键。为此在实时数据服务（如在线支付、跨行转账、反欺诈服务等）中，需要做到实时脱敏处理。智能数据脱敏技术应具备低延迟性，确保不影响服务的实时响应。此外，随着可被采集的用户信息日益多样化，金融信息的数据结构也趋向复杂化。这就要求脱敏技术

具备一定的柔性，在快速处理结构化数据的基础上，能够灵活处理半结构化和非结构化的数据。

（二）数据流通安全要求

数据流通的安全强调在数据共享、交易和传输过程中保护数据的安全性和隐私性。涉及金融数据在不同环境、系统甚至机构之间的传递。

1.跨环境、系统数据共享的“分级信任”策略

金融系统常存在生产、开发和测试等多套环境。生产环境直接服务客户，进行实际业务操作。开发、测试环境用于新功能开发以及单元、集成测试，以防止功能上线时出现问题产生实际影响。数据验证环境高度模拟生产的环境，用于监管报送等重要金融场景，使用仿真数据以确保测试结果准确性。此时应遵守跨环境的“分级信任”策略，根据请求环境、用户的重要性等级，对不同访问申请进行多层审批和多次身份验证和权限校验，确保数据在不同环境间传递时的安全，防止数据未经授权的访问。

由于金融机构业务范围广泛，客户数据通常被多维度地存储在不同的系统中。为构建准确的用户画像，数据要在多个系统和部门之间流动，甚至涉及到对外合作伙伴之间的数据共享。敏感数据在流通时，可能因安全漏洞等问题发生数据泄密。可以将数据存储在一个集中管理的数据平台，传递时遵循“分级信任”策略，根据请求系统的重要性提供差异化的信息展示。例如，给核心业务系统提供的业务数据保留更高的精度，而在外部系统中，

仅展示脱敏后的结果。另外，可根据数据的重要程度，通过流量限制控制大规模敏感数据的传输，以保证数据安全。

对于跨组织的数据流通，脱敏处理的方式要更加严格，如采用不可逆加密方式，保证数据即使被截获也无法被利用。另外，为了防止由于外部调用导致的敏感数据外泄，基于 API 进行数据交换的场景中可加入对数据接口的管控。

2.跨地区、跨境数据传输的合规性

跨境数据流动在金融机构全球化发展的今天变得更加普遍。这要求数据在跨境传输过程中能够满足不同国家和地区的数据保护法规（例如 GDPR、CCPA、PIPEDA[6]）。为降低违反不同地区法规的风险，智能脱敏技术应使脱敏后的数据不具备个人可识别性。如客户交易数据在跨地区、跨境传输前，剔除其中能够直接识别个人身份的信息，如姓名、身份号码等，并加密其他隐私信息，以减少因侵害客户隐私而导致违反当地法律法规的风险。

3.动态脱敏与数据可跟踪性

客户数据在流通过程中，信息被频繁地访问、传输和修改，要求数据脱敏时应有动态加工的能力。脱敏策略需要根据具体的业务场景，在数据从一个部门流向另一个部门的过程中调整。不同的数据用户，根据其权限和需求，获取的敏感信息应有所区别。比如客户数据、交易数据是为客户经理展示的非脱敏信息，而与信贷风险相关的信息则是风控部门访问的非脱敏信息。为此，智

[6] PIPEDA，全称《个人信息保护和电子文件法》(Personal Information Protection and Electronic Documents Act)，主要管理加拿大涉及商业活动的组织，特别是涉及跨省边界或国际的个人数据的传输。

能数据脱敏技术应具备根据不同访问用户对数据进行实时脱敏的能力。

同时，数据流通过程中的可追踪性也很重要，为确保数据在各个节点上的运行都可以被追溯，数据的传输路径应是可记录的。此外，数据跟踪能够提高数据的可信度和质量，在问题发生时能够清晰责任归属，优化业务流程。

（三）数据管理安全要求

数据存储划分为在线区、近线区和离线区数据[7]，以达到兼顾不同数据服务场景访问效率和经济化数据使用成本的目标。本节主要聚焦于近线区和离线区数据在存储、归档和销毁时的管理要求。

1.数据存储中的去标识化处理

数据存储中的去标识化处理主要应用于数据库中的近线区数据，该部分数据通常趋于静态存储，不再频繁更新或访问，这部分数据也是数据管理过程中最容易造成大规模数据外泄的风险点。智能数据脱敏技术需对存储的敏感数据进行去识别化或加密处理，以防止敏感信息在存储过程中泄露。例如，将客户的身份信息与交易数据加密后分开存储，确保不能直接识别客户的完整信息，即使单个数据集泄露，也无法识别客户的完整信息。

另外，大型金融机构通常会建立双活环境、灾备环境以保障发生故障时系统的稳定运行。但备份数据存储的安全级别通常相

[7] 交通银行数据中台根据数据的保留时长将数据划分为在线区、近线区和离线区数据，分别对应保留 2 年内历史数据、保留 2-7 年内的历史数据、超过 7 年的历史数据。

对较低，因此可以对灾备环境的数据进行智能数据脱敏处理，以进一步降低数据泄露的风险。

2.数据归档、销毁前的分级管理

为存储不断增加的客户数据，数据集群需要不断花费成本扩容节点，而扩容后的数据重分布、批量暂停也对服务的稳定提供造成压力。对于超出常规存储时间范围的离线区数据，可根据《2024 金融数据安全治理白皮书》[8]建议将金融数据分为核心数据、重要数据和一般数据三大级别，以确定数据是否应归档保存。而智能数据脱敏技术应能根据数据的不同级别，提供灵活且有效的脱敏策略。

而对于需要归档的核心级敏感数据，可在归档前进行多重脱敏（如数据遮盖、字符替换、哈希处理等），即使归档数据的某一层保护措施被攻破，剩余的脱敏方式仍然可以有效地保护数据安全。保证攻击者即便获取到部分已归档数据，也难以复原原始数据。

对于确认不再使用的离线区数据，应通过不可逆的脱敏处理，使数据变得没有意义，从而确保客户隐私及金融安全不受威胁，防止已销毁数据被恶意恢复。

3.管理过程中的风险评估与合规审核

在数据管理过程中，金融机构常要面对持续不断的行业风险评估和合规性审计。为此，在进行数据脱敏后，应保留详细的脱

[8] 《2024 金融数据安全治理白皮书》由中电金信发布，涵盖了金融数据安全治理的多方面内容。

敏日志、对应数据的加载策略以及生命周期转储策略的记录，以确保脱敏数据能达到国内外最新数据保护条例的要求。最后，智能数据脱敏技术应具备支持合规审核的功能，帮助金融机构及时发现数据管理中违反最新数据保护条例的潜在风险，以及进行调整脱敏策略后的安全风险评估。

北京金融科技产业联盟 BEIJING FINTECH INDUSTRY ALLIANCE

四、关键技术

（一）敏感数据识别

在金融行业，随着数据泄露、数据滥用事件的频发，如何有效识别和保护敏感数据已成为金融机构面临的一项重大挑战。本节主要讨论敏感数据识别的关键技术，涉及数据分类、模式识别、机器学习在金融行业的应用。

1. 数据分类

敏感数据识别前需进行数据分类，通常包含以下几个阶段：

（1）数据筛选

对组织的数据资产进行全面梳理，包括以物理或电子形式记录的结构化和非结构化数据，通过筛选、审核，识别其中的敏感信息，明确数据资产。具体可通过大数据平台的数据仓库、数据湖内登记的元数据进行全面扫描，初步筛选、生成数据使用情况报告。

（2）数据标记

可在根据数据敏感程度分类后，依托数据治理平台的标签系统对数据进行标识，该过程利用工具进行历史相同信息的自动化标记。其他可疑项可与安全部门、业务部门、数据部门等部门一起，在平台上对数据资产分类结果进行评审和完善，形成新的资产分类清单，并更新自动化标记工具，周期性迭代维护。

（3）数据分类

按照国家、行业、金融客户的数据分类保护要求，提取涉及核心数据、重要数据、个人信息的敏感数据范围，并调整数据资产的分类标记结果。同时，针对该部分敏感数据建立数据收集、存储、传输、使用、加工、导出、清除等全流程数据处理活动的分类保护措施。

2.模式识别

模式识别技术能有效从大量数据中自动鉴别敏感信息要素。常用的模式识别方法包括:

（1）入库识别

正则表达式常用于识别特定格式数据,如信用卡号、身份证、手机号、电子邮件等。处理中文等多字节复杂数据时，需结合Unicode和高阶正则表达式。对于姓名、地址等数据，则需利用特征库和自定义函数进行规则匹配，以识别数据分类。

（2）库内挖掘

对数据集内部存储的记录进行敏感信息的发现、提取。数据挖掘涉及数据清洗、数据转换、数据矫正、数据集成、数据挖掘算法的选择和开发、数据挖掘模型的构建和评估等多个环节。数据挖掘的主要目标是发现数据中隐藏模式、关系和规律，从而识别数据流转过程中产生新的敏感数据，并将其标识出来。

（3）出库兜底

数据在离开相关数据平台时，通过模式匹配算法，针对未标识敏感的数据列进行实时监控，避免中途链路加工过程中产生或遗漏的敏感数据被暴露到外部，在出库那一刻实时复核，该步骤技术作为检测数据泄露事件最后一环兜底。常规可利用数据定义类型及长度，快速缩小实时检测范围。

3.机器学习

引入机器学习技术，可以显著提升敏感数据识别的智能化水平。通过机器学习技术与数据分类分级规则、实际脱敏策略及规则的深度融合，能够实现自动化实时敏感数据发现、智能规则匹配等高效数据脱敏能力。这一技术不仅能够精准识别复杂场景中的敏感信息，还能根据数据特征自动优化脱敏策略，进一步提升处理效率与准确性。同时，系统支持分布式等多种部署方式，并具备自动化调优能力，能够灵活适应不同业务场景的需求，为金融机构提供更加安全、高效、智能的数据脱敏解决方案，助力其在数据安全与业务创新之间实现更好的平衡。

（1）自然语言处理

在处理文字或非结构化文本时，自然语言处理（Natural Language Processing，NLP）技术能够识别并标记文本中的敏感词汇。通过文本分析和分词，NLP 能够提取文本中的关键词并识别敏感信息，例如在转账备注中自动识别涉及金额、个人信息如身份证号和电话号码等敏感内容。NLP 中的命名实体识别（NER）

技术在此过程中尤为重要，它能有效识别文档中的人名、地点等实体信息，对于金融交易记录中的敏感信息识别尤为关键。

对于 NLP，除了能够提取敏感信息之外，通过语义分析的能力提升敏感数据的识别率也至关重要。比如基于理解的分词方法，通过让计算机模拟人类对语句的理解，达到识别词的效果，基于词向量的特征提取模型，通过向量的相似度来表示语义和语法相似度，另外基于情感词典的方法，挖掘正面、负面的情感分类。利用上述技术，NLP 处理技术能够让计算机更好地理解和分析人类语言的复杂性和多义性，从而提高人机交互的智能性和对敏感数据识别的识别率。

（2）监督学习

通过训练标注过的数据集，构建模型来自动识别敏感数据，常用的算法包括支持向量机（SVM）、决策树和随机森林等。

（3）无监督学习

针对没有标注数据的情况下，可以通过聚类分析等方法，识别出可能的敏感数据。该技术主要在处理新数据时，用于发现潜在的敏感信息。

（4）深度学习

深度学习利用具有自动特征提取能力的多层神经网络对图像、语音等数据进行分类和识别，具有强大的影像资源和复杂文本处理能力。

在敏感数据识别的实际应用中，除了识别的准确性，还需要考虑敏感信息识别的效率和精度。效率是指数据识别的速度和资源消耗，精度是指数据识别的准确率。为了提高效率，可以采用分布式计算、高效的脱敏算法、负载均衡等技术对算法进行优化；为了提高精度，可以通过大量数据标记和训练，以适应各种新数据和变种的敏感数据，采用集成学习等技术对多个模型的预测结果进行融合。相比传统的模式识别方案，机器学习相关技术可使得识别过程更加高效、智能化、精准化。

4.典型应用

针对金融行业的大数据平台建设，可通过元数据初筛、数据内容逐层复核，并结合上下游链路分析快速识别敏感信息，同时评估选择常用的识别方式。

（1）确定敏感数据范围

在大数据平台中，敏感数据通常指个人隐私信息，不包括国家安全层面的核心机密。直接标识个人的数据包括姓名、身份证号、邮箱、手机号、银行卡号等，而性别、生日、地址等则为间接标识。敏感数据特指那些一旦泄露或被非法使用，可能威胁个人安全的数据。因此，实践中重点关注姓名、身份证号、邮箱、手机号和地址等信息的保护，其他信息则不作为脱敏的重点。

（2）标注识别元数据信息

在识别敏感数据时，可以排除非整型数值、时间、日期、二进制等数据类型。整型数值中，小于电话号码位数的也可以排除。

对于字符型字段，一位字符类型可以快速排除。对于已确定的敏感数据（姓名、证件、电子邮件、电话号码、地址），可以通过数据治理平台或企业级数据资产平台快速筛选。不确定或潜在的敏感数据，则需依赖内容扫描技术。

（3）扫描敏感字段内容

利用元数据标识初筛后待扫描的数据字段，使用模式识别、机器学习等相关技术扫描实际内容，根据敏感特征发现对应字段的敏感属性。

若内容扫描代价过高，可通过随机提取一定比例数据进行判别，用于划分敏感数据类型；若扫描代价可接受，建议全字段内容扫描，以获得最佳识别效果。

字段内容识别，仅限平台数据录入侧的鉴别，不建议全平台全量扫描，即元数据标注、字段内容扫描适用约束在少量数据集判断。对于二次加工后产生的数据，即依赖加工链路的追踪，从敏感数据传递过来的数据保持其敏感性。

（4）追踪依赖链路

根据血缘分析能力或数据加工链路实时判定，提取字段链路的上下游关系，再依托源头已经标识敏感属性，推导下游依赖字段的敏感属性。从严而言，由敏感信息加工而来的数据均为敏感数据，特别是考虑现实场景，用户将敏感信息拆解成多个字段同时存储于一张表，丧失已有敏感特征，但其确定存在脱敏泄露风险，即严格意义上敏感数据加工需进行脱敏。当然现实生活中，

聚类、截选已大幅破坏数据的敏感属性，为了数据可流通性，可适当根据实现场景，标识特定场景下敏感属性的阻断。

总之，敏感数据识别是金融行业保障数据安全的重要环节。通过数据分类、模式识别、机器学习等关键技术的应用，金融机构能够有效识别和管理敏感数据。面对日新月异的技术迭代，特别是 AI 新兴技术发展，金融机构需持续更新其数据管理策略和技术手段，以确保数据安全与合规性。

（二）数据脱敏规则配置

1.脱敏策略选择

数据脱敏策略决定了如何处理和保护敏感数据。常见的脱敏策略包括字符掩码、数据加密、随机化和置换等。不同类型的数据适用不同的脱敏方法。例如，对客户的身份证号，可以使用部分掩码策略（如显示前六位和后四位，中间部分用*替代）；对交易金额，可以使用固定值替换策略，使其统一为固定值等。

2.配置方式选择

选择数据脱敏规则的配置方式需要明确以下几个方面的内容：需要脱敏的数据字段、选择的脱敏方法、脱敏级别和适用范围。数据脱敏规则的配置方式多种多样，可以根据不同的业务需求和数据特性进行灵活配置。以下是几种常见的配置方式：

（1）基于列的配置方式

基于列的配置方式是指对数据库中的特定列进行脱敏处理。例如，金融机构可以对客户表中的身份证号进行掩码处理，将其

替换为部分隐藏的形式（如“1234****5678”）。这种方式简单直接，易于实现，适用于需要对具体字段进行精确控制的场景。

（2）基于标签的配置方式

基于标签的配置方式是指根据数据的标签或元数据进行脱敏处理。例如，可以为敏感数据打上“机密”标签，然后对所有标记为“机密”的数据进行统一脱敏处理。这种方式具有较高的灵活性，能够动态调整和管理不同类别的数据脱敏规则，适用于大型复杂数据集的管理。

（3）基于角色的配置方式

基于角色的配置方式是指根据用户的角色和权限配置不同的脱敏规则。例如，普通用户只能看到脱敏后的数据，而具有高级权限的用户则可以访问原始数据。这种方式通过严格的权限管理确保数据的安全性和合规性，适用于需要区分数据访问权限的场景。

（4）基于条件的配置方式

基于条件的配置方式是指根据特定的条件对数据进行脱敏处理。例如，可以根据数据值的特定范围进行脱敏，或者只对满足某些条件的记录进行脱敏。这样的配置方式更具灵活性，能够根据具体业务需求进行调整，适用于需要精细化控制的场景。

（5）基于模板的配置方式

基于模板的配置方式是指使用预定义的脱敏模板对数据进行统一脱敏处理。例如，可以创建一个模板，对所有信用卡号统

一进行中间八位替换为星号的处理。这种方式能够简化脱敏规则的管理和维护，适用于需要批量处理的场景。

这些配置方式各有优缺点，选择时应根据具体的应用场景和数据保护需求进行综合考虑。通过合理配置数据脱敏规则，数据库管理员能够有效保护敏感信息，确保数据安全和隐私。

3. 规则实施

将配置好的脱敏规则应用到数据库中是数据脱敏的重要步骤。通过 SQL 脚本或配置文件，将脱敏规则加载到数据库中，并自动对相关数据进行脱敏处理。例如，可以编写 SQL 脚本，将客户表中的身份证号字段进行部分掩码处理。这样，无论是数据查询还是导出，用户看到的都是经过脱敏处理后的数据。

4. 规则验证与测试

在实际应用前，数据脱敏规则需要经过充分的验证与测试。测试的目的是确保脱敏后的数据在满足隐私保护要求的同时，仍然保持其业务可用性。这包括检查脱敏数据是否能够支持正常的业务操作和数据分析。基于测试环境，模拟各种业务场景，验证数据脱敏规则的有效性和可靠性。测试过程中，可以对比脱敏前后的查询数据，确保脱敏规则的正确实施。

5. 审计与监控

（1）脱敏规则的长期有效性

为了确保数据脱敏规则的长期有效性，需要建立完善的审计与监控机制。通过支持详细的日志记录和审计功能，可以帮助金

融机构定期检查和评估数据脱敏规则的执行情况。通过监控数据访问和处理过程，及时发现和解决潜在问题。例如，可以定期生成审计报告，检查是否有未脱敏的数据泄露，确保数据脱敏的安全性和合规性。

（2）脱敏过程的可追溯性

基于安全性维度的考量，脱敏过程应当具有可追溯性，即需要记录和跟踪每一个脱敏操作以便在需要时进行审计。可追溯性对于金融机构尤为重要，因为它不仅能够帮助确认是否正确执行了脱敏操作，还能查找可能出现的问题。通过在数据脱敏过程中提供详细的日志记录功能，从而记录每次数据脱敏的具体时间、操作人员和具体操作内容。这些记录能够在后续审计中提供关键证据，确保脱敏操作的透明度和脱敏过程的可追溯性。

（三）数据脱敏可算不可见引擎

数据脱敏可算不可见引擎是数据脱敏技术的核心组件，旨在实现数据的隐私保护和可计算性。它不仅要保证数据在使用过程中不暴露敏感信息，还要确保数据的计算和分析能力不受影响。

引擎通过隐私保护计算、动态数据脱敏和数据虚拟化技术，实现数据的安全和高效处理。可以在对指定数据按照规则进行脱敏处理的基础上，保留数据的关联关系。具体来说，可算不可见引擎在数据库内核中采用脱敏前的原始数据进行关联运算，而在将数据发往库外时将数据做脱敏处理，确保了敏感数据在数据库内可以参与运算而在数据库外不可查看原始数据的目的。以身份

证号为例，在数据库内，引擎可以通过未脱敏的身份证号过滤数据，关联用户信息表和交易流水表得到准确结果。针对待出库的身份证号，引擎会将其划分为地址码（前 6 位），出生日期码（中间 8 位），顺序码（3 位）和校验码（最后 1 位）。行政区划代码根据真实的行政区划生成权重值，并到系统内置行政区划库中进行脱敏计算，形成新的代码。出生年月部分，系统会根据指定偏移值进行偏移值计算，生成新的出生年月。三位系统顺序码按照纯数字脱敏方式脱敏，并根据上述计算结果生成校验位。最终组合成脱敏后的数据出库。达到库内数据正常运算，出库数据不可见的效果。可算不可见引擎主要涉及如下技术。

1.隐私保护计算

隐私保护计算技术在可算不可见中起到了重要作用，通过对隐私数据先运算再加密，达到在不泄露敏感数据的情况下，对数据进行处理和分析。借助隐私保护计算技术，原始敏感数据可以在数据库内参与运算，仅在出库时刻（返回结果时）才会做脱敏处理，确保了金融敏感数据的可用性。

2.动态数据脱敏

动态数据脱敏是数据脱敏可算不可见引擎的另一项重要技术，通过在数据访问和处理的过程中，实时进行脱敏处理，可以为不同角色、不同权限、不同数据类型执行不同的脱敏方案，从而确保返回的数据可用而安全。摒弃业务应用层脱敏依赖性高、代价大等痛点，将数据脱敏功能内置到数据库产品自身的安全能

力中，使数据脱敏解决方案具备完整、安全、灵活、透明、友好的特点。如下图所示，动态数据脱敏引擎是基于原有数仓底座的新引擎，在用户交互界面下，与 SQL 引擎和存储引擎直接交互。在 SQL 语句执行过程中，随着查询解析、重写操作实时触发的脱敏行为，SQL 引擎根据重新构建的 Query Tree 生成最优的执行计划，使得对象变化时表、视图、存储过程均可实时脱敏，充分利用分布式框架、SQL 引擎保证动态脱敏性能，易用性、扩展性、维护性更好。这样，用户在查询和分析数据时，看到的始终是经过脱敏处理后的数据，从而降低了数据泄露的风险。

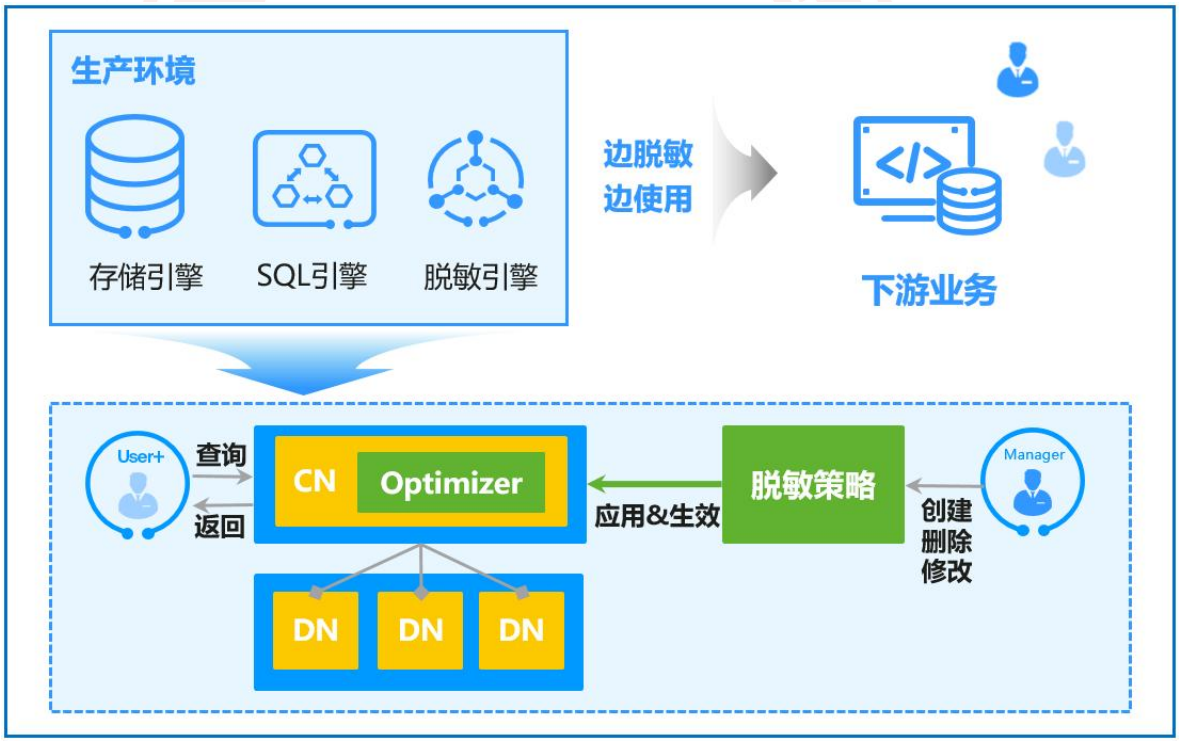

图 1 动态数据脱敏引擎执行流程图

动态数据脱敏引擎提供了灵活的动态脱敏配置功能，可以根据业务需求，实时调整脱敏策略和规则。动态数据脱敏不仅提升

了数据安全性，还增强了数据的实时性和可用性。与传统脱敏技术相比，动态数据脱敏存在以下技术优势:

（1）良好的底座协同性

动态脱敏引擎贯穿于数仓底座的诸多环节，基于预先配置好的脱敏策略，参与 SQL 引擎每条语句的解析、重写、优化与执行。得益于原厂优势，动态脱敏引擎更贴近底座 SQL 引擎本身，基于策略的重写逻辑直接作用于内核优化器关键信息载体 QueryTree 而非原始 SQL 语句，同时，表对象所关联的脱敏策略信息是直接与集群元数据实时获取的，从而使得视图、含有动态 SQL 的存储过程、UDF 函数、多层嵌套、多源数据、多次 INSERT 临时表等复杂业务场景也可以在解析执行过程实时动态地屏蔽敏感数据。

（2）脱敏过程用户无感知

动态脱敏功能形成了一套完备、友好、易用的使用框架，提供灵活可配置的脱敏策略语法，允许用户指定数据脱敏的触发条件、脱敏字段及脱敏效果函数集合，策略也允许适时关闭或开启。当查询语句执行过程中满足触发条件（通过当前用户角色界定）时，会自动生效预置的脱敏效果，在查询真正对外暴露执行结果的那一刻予以展示，从而做到脱敏过程无感知。

（3）灵活可扩展的脱敏策略

为了既达到脱敏的目的又达到保留数据特征的目的，通常对于不同类型的文本数据，期望展示不一样的脱敏效果。以姓名和邮箱为例，姓名通常期望脱敏成“张**”，而邮箱通常期望脱敏

成“***@163.com”，从而使得脱敏效果依然具有原始数据本身的可分辨性。动态脱敏引擎支持定制化脱敏效果，客户可结合自身业务场景识别敏感数据并对业务表的指定列灵活预置脱敏策略。

3.数据虚拟化

数据脱敏可算不可见引擎利用数据虚拟化技术，将脱敏数据与原始数据分离，并通过访问控制策略，确保只有经过授权的用户才能访问脱敏前的原始敏感数据。通过支持多种数据虚拟化方案，数据脱敏技术可以根据金融机构的需求，灵活配置和管理数据访问和脱敏策略。数据虚拟化技术不仅提高了数据安全性，还简化了数据管理和访问控制的复杂性。

（四）数据脱敏核心算法

数据脱敏核心算法是数据脱敏技术的基石，直接影响到数据脱敏的效果和安全性。通过集成多种数据脱敏核心算法，数据脱敏技术为金融机构提供了灵活、可靠的数据脱敏方案，确保数据在保护隐私的同时，仍具备业务可用性。常见的脱敏算法包括哈希脱敏算法、遮盖脱敏算法、替换脱敏算法、变换脱敏算法和洗牌脱敏算法等，表 1 展示了这几种常见脱敏算法的脱敏效果。

表 1 常见数据脱敏核心算法的脱敏效果

算法名称	敏感数据（以手机号码为例）	脱敏处理后的敏感数据
哈希脱敏	12345678901	8da28dd7ed4357331a0d05202acb5b6bc7be4b7530e24e1cadb1086d4deb7ce5
遮盖脱敏	12345678901	123******01
替换脱敏	12345678901	12371290301

变换脱敏	12345678901	[12000000000,13000000000]
洗牌脱敏	12345678901	23415678901

1.哈希脱敏算法

哈希脱敏算法是一种常用的数据脱敏技术，通过将敏感数据使用哈希函数转换为固定长度的字符串，使得原始数据无法被直接识别或逆向还原。哈希函数是一种单向函数，即数据一旦经过哈希处理，就无法轻易逆向解密还原到原始数据。因此，哈希脱敏算法在保护数据隐私方面具有较高的安全性。

在实际应用中，哈希脱敏算法可以用于保存各种敏感信息，如信用卡号、社保号码、电子邮件地址等。例如，将客户的身份证号码通过 SHA-256 哈希函数进行处理，生成的哈希值可以用作替代标识符，从而隐藏原始身份证号码的真实信息。

哈希脱敏算法的优势在于其计算效率高、易于实现且能够提供较强的隐私保护。然而，由于哈希函数是确定性的，即相同的输入总是会产生相同的输出，因此可能会遭遇字典攻击或暴力破解。在实际应用中，常常结合使用 Salt 技术，即在输入数据前附加随机数，以进一步增强哈希脱敏的安全性。

总的来说，哈希脱敏算法在数据脱敏中具有重要作用，能够有效保护敏感数据的隐私和安全。

2.遮盖脱敏算法

遮盖脱敏算法是一种常见的数据脱敏技术，通过替换或隐藏敏感数据中的部分或全部字符，使其在数据处理中不暴露原始信

息。遮盖脱敏的基本原理是将敏感数据用特定的符号(如星号“*”或问号“？”)替换。例如，将信用卡号“1234-5678-9012-3456”掩码为“1234-****-****-3456”，或将身份证号码“123456789012345678”掩码为“1234********5678”。这样，脱敏后的数据仍然保持了原始数据的格式和长度，但敏感信息部分被遮盖。

遮盖脱敏算法可以应用于多种场景，在软件开发和测试过程中，使用遮盖脱敏技术可以生成与生产环境一致但不包含真实敏感信息的数据，用于测试和调试，避免数据泄露风险。在数据分析和报表生成过程中，使用遮盖脱敏技术可以保护个人隐私，同时保留数据的统计和分析价值。在与客户交互过程中，使用遮盖脱敏技术可以在客户信息展示时隐藏部分敏感数据，提高数据安全性。

遮盖脱敏通过隐藏敏感数据，能够有效保护个人隐私和敏感信息，防止未经授权的访问和泄露。此外，遮盖脱敏技术能够保持数据的原始格式和长度，使得脱敏后的数据在系统和应用中仍然可以正常使用，而不需要对现有系统进行大规模修改。遮盖脱敏算法实现相对简单，易于在数据库层面或应用程序层面实现和部署，对现有系统的影响较小。通过对敏感数据进行遮盖处理，能够有效降低数据泄露的风险。即使在数据泄露事件中，攻击者也无法获取完整的敏感信息，从而提高了数据的安全性。但在需

要对原始数据进行复杂分析时，遮盖脱敏算法可能会影响数据的部分准确性，需要配合可算不可见功能使用。

3.替换脱敏算法

替换脱敏算法通过将原始敏感数据替换为伪造数据，使处理后的数据无法被识别到原始信息。这种方法确保数据结构和格式不变，但不包含任何真实的敏感信息。替换脱敏适用于测试、数据分析等场景，能够保护个人隐私和敏感信息。

具体来说，替换脱敏可以通过静态替换、动态替换和部分替换等方式实现。例如，将客户姓名替换为随机生成的名字，或将手机号替换为符合格式的随机号码。替换后的数据保留了与原始数据相同的结构和格式，但不包含真实的敏感信息。

替换脱敏的优点在于保护隐私，保持数据格式和结构，易于实现和部署。替换后的数据可以用于开发和测试环境，避免泄露真实数据。然而，生成的伪造数据可能缺乏真实性，因此在实际应用中需注意选择适当的替换数据。

4.变换脱敏算法

变换脱敏算法通过对数据进行变换处理，使原始数据无法被识别。例如，将具体年龄变换为年龄段（如 20-30 岁），或将精确的地址模糊化为城市级别。变换脱敏在不影响数据分析的前提下保护隐私，适用于统计分析和数据挖掘。

具体来说，变换脱敏可以通过泛化、加噪等方式实现。例如，将详细的出生日期转换为年份或年龄段，或在数据中加入随机噪

声，使得原始数据难以被推断。变换后的数据在保护隐私的同时，仍然保留了数据的统计特性。

变换脱敏的优点在于灵活性强，能够保留数据的统计特性和分析价值，适用于大数据分析和挖掘。通过合理设置变换规则，可以在保护隐私的同时，保证数据分析的准确性。然而，变换脱敏的效果依赖于变换规则的选择和实现，需要结合具体业务需求进行合理设置。

5. 洗牌脱敏算法

洗牌脱敏算法是一种有效的数据脱敏技术，通过随机打乱数据中的记录顺序或字段顺序，使其在数据处理中无法直接关联到原始信息。这种方法在保护敏感数据隐私的同时，保留了数据的统计特性。

洗牌脱敏算法的基本原理是利用随机化技术对数据进行重新排列。例如，将客户列表中的记录顺序随机打乱，使得原始记录无法与打乱后的记录一一对应；或者对数据表中的某些字段进行重新排列，使得脱敏后的数据无法恢复到原始状态。通过这种方式，洗牌脱敏能够有效隐藏数据之间的关联性，防止未经授权的访问和滥用。在需要共享数据但又不能暴露敏感信息的场景中，使用洗牌脱敏技术可以确保数据在传输和交换过程中的安全性。

洗牌脱敏通过随机打乱数据的顺序或字段，能够有效保护个人隐私和敏感信息，防止未经授权的访问和泄露。而在保护隐私的同时，洗牌脱敏保留了数据的整体统计特性和分析价值，使得

脱敏后的数据仍然可用于统计分析和研究。通过对敏感数据进行随机化处理，洗牌脱敏能够有效降低数据泄露的风险。即使在数据泄露事件中，攻击者也难以通过打乱后的数据恢复到原始数据，从而提高了数据的安全性。而且洗牌脱敏技术可以根据业务需求灵活配置打乱规则，适用于各种不同的数据类型和应用场景。

综上所述，哈希脱敏、遮盖脱敏、替换脱敏、变换脱敏和洗牌脱敏算法各有特点和适用场景。哈希脱敏通过哈希函数保护数据隐私，适用于需要唯一标识的场景；遮盖脱敏通过隐藏部分字符保护数据，适用于开发和测试环境；替换脱敏通过伪造数据替换原始信息，适用于数据分析和处理；变换脱敏通过变换数据保护隐私，适用于统计分析和数据挖掘；洗牌脱敏通过随机打乱数据顺序保护隐私，适用于数据分析和数据共享。在实际应用中需要结合具体业务需求和数据保护要求选择合适的脱敏算法。

五、展望建议

（一）持续探索研究，加强数据识别和脱敏技术性能优化

由于数据类型多样、算法数量线性化增长等突出问题，实时动态场景下的数据识别面临的主要技术难点是对识别算法的性能要求极高。如何在不影响业务应用的前提下，实现快速准确的资产识别与实时动态脱敏是亟需解决的问题。首先，需要对业务的具体需求和数据资产特征进行充分调研，比如有的数据资产识别具有较强的规则性，有的数据资产识别需要多种模式结合才能表达出资产的正确组成结构；其次，根据不同的业务需求和数据资产特征，选择合适的识别算法，均衡业务性能与数据识别的需求。

针对非结构化数据脱敏，需要通过研究敏感文本识别、敏感图像识别的算法，提升算法类型覆盖、算法召回率，提升脱敏效果。比如针对文本数据识别需要尽可能恢复文本结构信息以增强对使用场景的感知能力，进而提升识别的准确率和运营能力。针对图像数据，需要在对图片进行预处理的基础上，使用目标检测算法与图片分类算法进行初步划分，并识别出图片的风格、渠道和拍摄环境等背景信息；对识别出类型的图片调用 OCR 算法提取文本信息；结合目标检测、图片分类、文本匹配和校验结果进行精细化的敏感图片数据分类分级。然后需要根据业务需求，灵活搭配，使针对敏感数据的脱敏能够满足数据原始属性、关联性、可追踪性以及准确性等要求。

（二）坚持守正创新，提升数据脱敏更加安全高效

随着人工智能技术广泛应用，如何使数据脱敏技术有效满足多模态数据交互流量的不断增长和复杂多变的安全处理业务场景急需解决。因此，需要将数据脱敏技术与人工智能的自主学习和强大的数据分析能力有机结合，通过定义敏感数据基本特征，利用样本进行训练学习，数据脱敏方法也可通过人工智能进行灵活选择，防止同样的数据进行同样的脱敏处理后可能带来的可链接攻击；人工智能通过对脱敏后数据进行实施监控，及时发现并纠正潜在的安全问题，实现数据脱敏技术在人工智能时代的灵活安全使用。此外，人工智能能够根据数据的使用场景和用户需求动态调整脱敏策略。这种自适应能力使脱敏过程更加灵活，以应对不断变化的安全需求。人工智能通过学习用户的行为模式，可以优化脱敏规则，减少人工干预，提升数据脱敏的准确性和效率的同时，兼顾数据安全和业务分析目标。

图计算作为一种高效的数据处理与分析工具，近年来在众多领域展现了显著的应用潜力。通过深度融合图计算技术，数据脱敏的效率与安全性得以进一步提升。首先，通过图结构，数据脱敏可以更好地理解和处理数据之间的关联性。例如，在金融领域，图计算可以帮助识别和分析不同账户之间的复杂交易关系，从而在脱敏过程中保持这些关系的完整性和安全性。其次，图计算可以用于分析数据特征，生成更智能的脱敏规则。例如，通过图算法分析数据节点之间的关系，可以自动识别哪些数据需要更严格

的脱敏措施，从而优化脱敏策略。在动态数据脱敏中，图计算可以实时分析数据流，动态调整脱敏策略。这种动态调整能力使得脱敏过程更加灵活，能够应对不断变化的数据访问需求和安全威胁。另外，图计算可以用于检测数据中的异常行为，例如异常的数据访问模式。通过分析图结构中的节点和边的变化，可以及时发现潜在的安全威胁，并采取相应的防护措施。

（三）强化标准指导，完善数据脱敏技术机制建设

数据脱敏技术在金融业数据治理和流通方面发挥着重要作用，通过数据脱敏，可以有效实现数据降级，使原本不能流通的高敏感数据转化为可以流通的低敏感数据，助力平衡数据安全与数据流通需求。但是目前针对不同等级和类型的数据如何有效应用数据脱敏技术，尚缺乏行业统一的标准指导。因此，一方面实现统筹规划，建立和完善数据识别、数据脱敏相关标准体系，以标准化手段指导数据脱敏技术应用工作的体系化建设和业务推进；二是制定数据识别、数据脱敏相关技术标准，推动数据脱敏关键技术、算法、指标的研究和应用，不断提升数据脱敏技术的改进，以标准促进技术创新；三是制定数据脱敏技术相关测评方法和应用指南，制定科学评估体系和应用建议，建立健全评估与监督机制，严格落实政策要求。

（四）完善基础设施，推进数据脱敏体系化应用

在业务快速增长、数据规模和复杂程度激增的背景下，数据安全治理面临风险感知后置、场景覆盖较为被动、网状数据链路

改造成本高等挑战。与此同时，数据脱敏等安全能力也需通过统一能力下发机制，实现整体安全防护水平的快速提升。为此，需完善网络安全基础设施建设。一方面，构建深入业务场景的数据采集体系，为安全策略制定、事件处置及数据流转链路刻画提供全面的场景化数据支持；另一方面，建立统一高效的安全组件注入与更新平台，使业务应用能够以极低成本快速接入数据脱敏等安全能力，同时支持安全能力的独立配置与动态升级，实现安全防护与业务发展的深度融合。

附录：金融业智能数据脱敏应用实践

案例一：邮储银行数据脱敏应用实践

1.案例背景

金融行业的业务生产系统积累了大量包括账户和客户隐私等敏感信息的数据。如果这些数据产生外泄，带来经济损失的同时，会给银行的声誉及社会效应带来负面影响。

随着未来监管的要求越来越高及金融机构内部对业务测试数据的质量的提高，金融机构数据调用的频率增加，用途逐渐多样化，其内部审计也提出了数据安全的相关要求。依据本行测试数据管理需求，防止金融机构重要数据资产泄露风险，依托数据脱敏系统的建设，建立金融机构数据安全保护体系。

2.建设目标

数据脱敏系统项目建设目标是建设适用场景广泛、内置丰富脱敏算法、敏感字段自动发现、支持多种数据库及文件结构脱敏、高效安全的脱敏平台。

搭建数据脱敏常态化机制，增强数据脱敏自动化、流程化水平，对接数据提取外发流程并提供文件脱敏入口，对外发数据文件进行脱敏，降低敏感数据泄密风险。

3.业务需求分析

通过数据脱敏系统，实现隐私数据脱敏（变形和保护）的自动化及可视化，保障数据安全，加强调用管理，提升操作效率，

满足审计及监管部门要求，有效防止生产数据中敏感信息的泄漏，保障数据安全，规避数据风险。

数据脱敏系统能完全保证隐私数据脱敏时的应用逻辑：例如，姓名使用随机但有效、唯一的姓名来替换，而不是使用无意义的文本字符串；证件类使用随机、有效、原始的证件类型（身份证对应身份证，军官证对应军官证等）；在使用技术屏蔽一些数据的同时，其他诸如银行代码和账号之类的数据也必须是虚拟的而且保持其在上下文中有效。数据脱敏系统支持灵活的配置方式（包括字段信息匹配、数据信息匹配）来自动探测数据库敏感信息字段。既能轻松找到敏感数据，又能防止纯人工操作引起的疏漏。

4.数据脱敏功能详细介绍

（1）脱敏方式

数据脱敏系统提供数据源的统一管理，可支持各种不同类型的数据源，以统一方式进行配置并获取访问。提供从多个同构、异构数据库中定制关联关系，以实现跨数据源的联邦关联抽取能力。

数据库类型包括：

- 支持关系型数据库脱敏，如Oracle、Sql Server、Informix、Mysql、DB2、Postgre sql、达梦、人大金仓；
- 支持大数据脱敏，如Teradata、Impala、Hive；支持txt,csv,dmp,Excel等文件类型作为脱敏源和目标，且支持远程

ftp和sftp发现;

（2）数据源与目标配置

数据脱敏系统支持从生产数据库通过系统生成导出文件装载至非生产数据库，数据导出格式非常灵活和广泛，对间接脱敏（从生产数据库或文件通过脱敏系统生成脱敏后数据的导出文件，用于装载至非生产数据库及文件）、直接脱敏（将生产数据库及文件不落地直接脱敏至非生产环境数据库及文件）两种脱敏方式均支持。

（3）数据脱敏流程

数据脱敏系统能自动将数据库中的可能敏感信息扫描并展示，同时提供敏感字段的样本数据，供管理员参考分析。

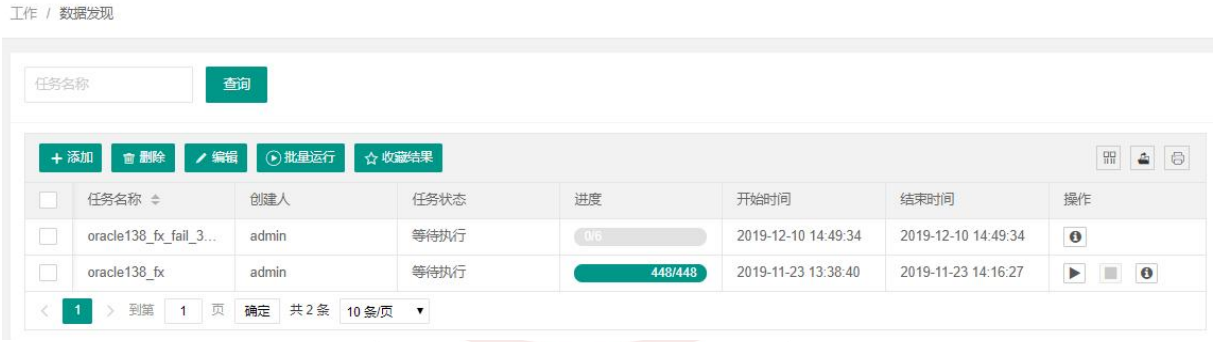

图 2 新建数据分析任务

数据处理任务即：数据脱敏，指对某些敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。

图 3 查看数据处理任务

（4）脱敏算法

数据脱敏系统提供了预定义的数据变形策略，默认支持常用中文、英文、电话号码、证件号码、中英文地址等屏蔽规则，内置了一些常用的算法，包括确定随机化、模糊化、置空、乱序排列、重复值屏蔽、随机替换、特定规则替换等算法。

数据脱敏系统除了上述内置的屏蔽规则之外，还支持添加基于DB和JAVA的自定义屏蔽规则，可以满足用户所有的屏蔽需求。

（5）数据处理预览

数据脱敏系统在数据脱敏处理之前，可通过数据处理预览，快速查看数据脱敏后的效果，可以真实地了解脱敏后的结果，方便用户使用前预判是否按照此算法进行脱敏，可降低因脱敏结果不合格造成脱敏反复的时间。

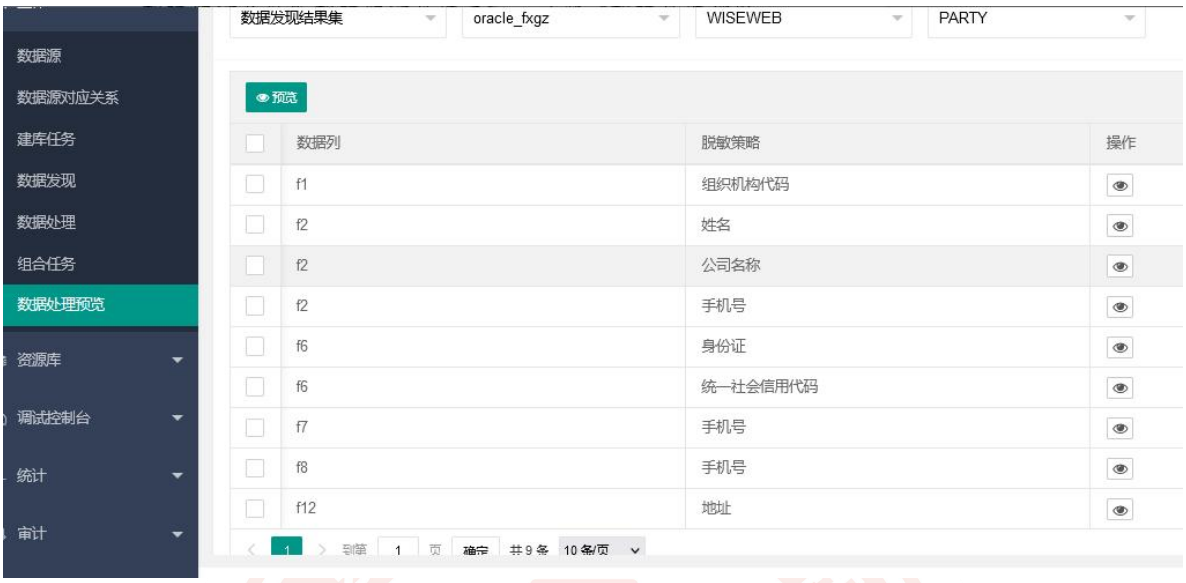

图 4 数据处理预览

（6）自动化脱敏

数据脱敏系统实现全自动化数据脱敏，通过状态返回值判断不同节点的执行情况，脱敏系统自动识别文件格式和文件路径遍历，实现自动任务创建及执行，完成脱敏流程自动化操作。

脱敏引擎采用spark/livy 组合，支持多主机多节点大规模并行，处理性能主要受脱敏运算节点物理计算能力，网络存储能力影响，大幅提高原有脱敏运行效率。

同时，通过引入Livy避免将大数据集群直接暴露给脱敏服务器，避免由于脱敏服务器直连导致的网络配置，无权限验证等问题。并且脱敏服务可灵活支持多套大数据集群。

案例二：蚂蚁集团数据脱敏应用实践

1. 案例背景

数据保护伞是蚂蚁集团多年的数据安全沉淀，为客户提供大数据安全管理能力， 基于数据资产嗅探、图计算、语言模型、行为关联分析等技术实现数据源管理、分类分级、数据资产、识别发现、脱敏水印、数据操作审计、风险管控、数据溯源分析等功能，保障数据从采集传输、储存处理、到交换共享的全生命周期的安全防护。产品整体架构如图5所示。

图5 数据保护伞整体架构

2. 建设目标

数据保护伞致力于提供动态脱敏能力和标准脱敏服务能力，根据保护伞对敏感信息的定义和脱敏策略的制定，智能识别系统展示内容中存在的敏感信息并进行脱敏，达到防止敏感信息泄露的目标，并能够将企业内部的脱敏水位进行统一管理，在保障安全水位的同时极大地提升了安全管理效率。

3.功能介绍

针对数据脱敏，数据保护伞具备以下能力:

支持多种脱敏场景根据不同项目、用户组进行自定义脱敏，并为每个脱敏场景制定不同敏感数据的脱敏规则;支持应用动态脱敏和运维动态脱敏两种脱敏形式对应的多种脱敏场景选择;

脱敏能力开放支持第三方系统或数据查询使用入口通过API方式调用脱敏服务，以供各类业务系统实现敏感数据脱敏，在数据脱敏管理页面配置脱敏场景和场景码信息，对接的业务系统在有用户查询敏感数据时调用脱敏服务，数据保护伞根据用户配置信息返回脱敏后的数据。

脱敏规则定义支持针对不同的场景、不同的敏感数据制定脱敏规则，脱敏方式包括有保留格式加密、HASH加密、掩盖、字符替换、区间变换、取整、置空等;支持针对每个脱敏规则设置白名单，白名单用户在进行相应场景数据访问时，不进行脱敏处理。

4.应用实例一

以下是数据保护伞在某银行的应用案例，如图6所示。

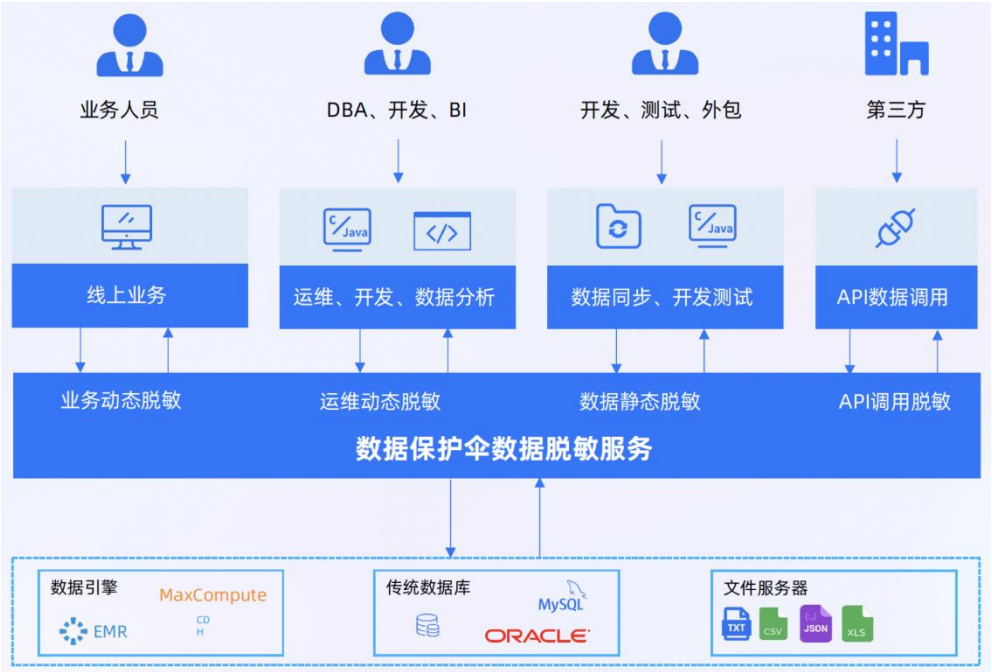

图6 蚂蚁数据保护伞在某银行的应用案例

（1）应用数据保护伞前

在应用数据保护伞之前，某银行存在如下问题:

- DBA等运维人员在运维过程中，可查看敏感数据，有数据泄露的风险;
- 应用系统无法改造，无法对数据进行脱敏后展示;
- 高权限账号共用、滥用;
- 数据获取入口多且复杂，无法进行统一管控。

（2）应用数据保护伞后

在应用数据保护伞之后，达到如下效果:

- 多种生产场景下的脱敏要求，保障用户的敏感数据不被泄漏的同时不影响正常业务;
- 业务系统老旧，升级困难，改造周期长成本高等问题得

到了很好地解决;

- 多类数据获取入口可统一管控;
- 根据用户权限细粒度管控脱敏数据。

5. 应用实例二

以下是数据保护伞在某金融机构的应用案例，如图 7 所示。

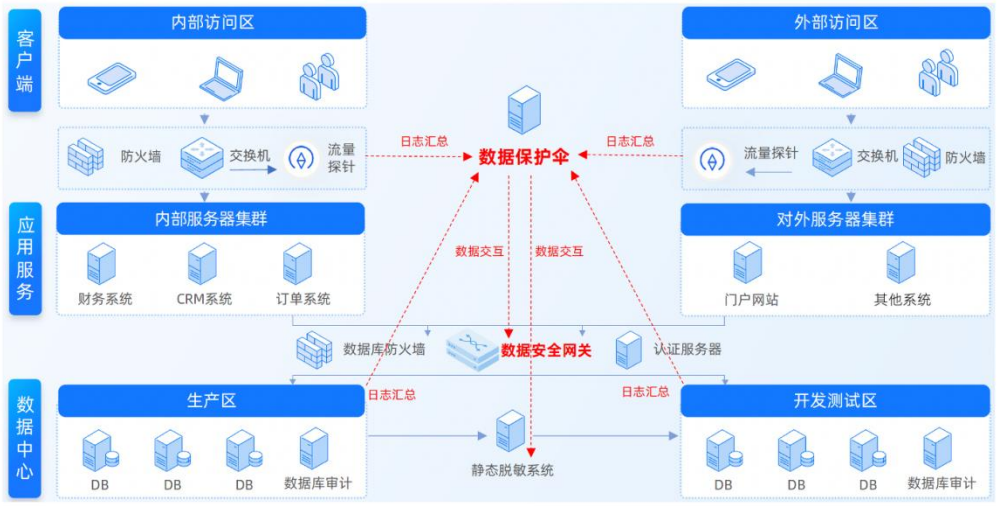

图7 蚂蚁数据保护伞在某金融机构的应用案例

（1）应用数据保护伞前

在应用数据保护伞之前，某金融机构存在如下问题:

- 内部人员违规访问内部系统存在数据泄露、滥用风险;
- 运维人员违规访问数据库、服务器，存在数据泄露风险;
- 第三方通过接口违规提取截留敏感数据;
- 日志分散，难以集中分析，人工分析成本大;
- 安全部门对业务部门使用敏感数据情况不了解，难以制定安全策略。

（2）应用数据保护伞后

在应用数据保护伞之后，达到如下效果:

- 多场景多类数据访问统一管控;
- 日志集中管控分析,多维关联分析引擎,分析条件丰富、灵活，适用于客户复杂的业务场景，事件分析定位更精准;
- 多种生产场景下的脱敏要求,保障敏感数据不被泄漏同时不影响正常业务;
- 可视化了解敏感数据使用、流转情况，帮助企业用户制定合理安全策略。