ICS 35.020
CCS L 70

T/HEBQIA

团　　体　　标　　准

T/HEBQIA XXXX—2025

数据资产整合与挖掘规范

Data asset integration and mining specifications

（征求意见稿）

2025－XX－XX 发布　　2025－XX－XX 实施

河北省质量信息协会　　发 布

目　　次

前　言

本文件按照 GB/T 1.1—2020《标准化工作导则　第 1 部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由联城科技（河北）股份有限公司提出。

本文件由河北省质量信息协会归口。

本文件起草单位：联城科技（河北）股份有限公司、唐山国控科创集团有限公司、唐山联信数据有限公司、XXXXX。

本文件主要起草人：高峡、高月仁、吕晓栓、席啸、杨海涛、刘紫群、赵欣宇、何岩、高帆、宋建旭、XXXXX。

数据资产整合与挖掘规范

1 范围

本文件规定了数据资产整合与挖掘的基本原则、参与主体要求、数据架构、技术架构、数据流转模式、数据源接入要求、数据资源整合流程、挖掘算法选型、安全管控要求。

本文件适用于政务服务、医疗健康、教育民生、城市规划等公共服务领域的数据挖掘活动。覆盖政府部门、企（事）业单位、第三方技术服务机构等相关主体。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据资产 data asset

合法拥有或者控制的，能进行计量的，为组织带来经济和社会价值的数据资源。

3.2

数据资产整合 data asset integration

将分散在不同来源、不同格式、不同结构的数据资产，通过采集、清洗、转换、关联、虚拟化等一系列技术手段和管理流程，形成逻辑上统一、物理上可灵活调度的数据集合，实现数据资产的统一访问、共享和管理的过程。

3.3

数据资产挖掘 data asset mining

从整合后的数据资产中，运用统计分析、机器学习、深度学习、隐私计算等技术方法，发现潜在规律、关联关系、趋势特征等有价值信息，并将其转化为业务决策支持、产品创新、服务优化等成果的过程。

3.4

洋葱加密 onion encryption

对数据资产的每个字段根据使用场景、使用方身份、使用价格等因素，采用不同的加密算法进行多层加密处理，形成类似“洋葱”的加密结构，实现数据资产的细粒度访问控制和使用控制，防止数据未授权拷贝、传输和泄露的多层加密技术。

4 基本原则

4.1 合规性

敏感数据处理应取得数据主体同意，跨部门数据流通履行安全评估程序，禁止未经授权的数据采集、使用及传输。

4.2 安全性

应采用“分层防护”策略，从数据接入、存储、计算到流转全环节实施安全管控，优先使用洋葱加密、隐私计算等技术，确保原始明文数据不出域，密文数据流转可追溯，防止数据泄露、篡改或滥用。

4.3 标准化

应统一数据接入接口、格式标准、融合规则及算法评估指标。

4.4 实用性

算法选型、模型设计应适配公共服务场景需求，避免过度复杂技术导致资源浪费或响应超时。

4.5 可追溯

通过区块链存证记录全流程操作日志，确保数据来源可查、使用可追、责任可究。日志应包含操作主体身份信息、授权范围、数据使用目的及结果去向，支持监管部门随时调阅审计。

5 参与主体要求

5.1 数据提供方

5.1.1 应具备合法的数据资产所有权或使用权，确保所提供的数据资产符合法律法规和相关政策要求，不得侵犯第三方合法权益，如数据主体的隐私权、知识产权等。

5.1.2 应部署数据客户端，通过客户端完成数据资产的注册、接入、协议制定和使用控制策略配置。

5.1.3 应制定数据资产的分类分级标准，根据数据资产的敏感程度、重要性、使用范围等因素，对数据资产进行分类（如公开数据、内部数据、敏感数据、机密数据）和分级，并针对不同类别和级别的数据资产制定相应的整合策略和安全保护措施。

5.1.4 应建立数据资产质量保障机制，在数据资产接入整合前，对数据的完整性、准确性、一致性、时效性、唯一性等质量指标进行检测和评估。对于不符合质量要求的数据资产，应进行清洗、修复或标注，确保整合后的数据资产质量满足后续挖掘和使用需求。

5.1.5 应按照数据使用协议的要求，对数据资产的使用范围、使用方式、使用期限、使用次数等进行明确界定，并通过客户端和云端的技术手段对数据资产的使用过程进行监控和追溯。当发现数据使用超出协议范围时，应具备数据冻结、销毁等控制能力。

5.2 数据消费方

5.2.1 应通过数据客户端接入空间云端，完成身份认证和授权申请，获取数据资产的使用权限。身份认证可采用多因素认证方式，确保身份的真实性和唯一性。

5.2.2 应遵守数据使用协议和相关法律法规，按照协议约定的用途、方式和范围使用数据资产，不得超出授权范围使用数据，不得将数据资产转让、出租、出借或用于协议约定以外的其他目的，不得泄露、篡改、破坏数据资产。

5.2.3 应具备相应的数据处理和存储能力，根据数据资产的类型和规模，配置符合要求的硬件设备和软件系统，确保数据资产在使用过程中的安全性和稳定性，如采用加密存储、访问控制、备份恢复等措施保护数据资产。

5.2.4 应建立数据使用日志记录机制，记录数据资产的查询、下载、计算、分析等使用操作。日志内容应包括操作时间、操作人、操作内容、数据标识、使用结果等信息，并定期将日志同步至空间云端的监控审计系统，配合监管和审计工作。

5.2.5 当数据使用完毕或数据使用协议到期时，应按照协议要求和客户端提示，对数据资产进行安全销毁，并向数据提供方和空间云端提交数据销毁证明，确保数据资产不被留存和滥用。

5.3 数据运营者

5.3.1 应搭建和维护空间云端平台，确保平台的稳定性、可用性和安全性。平台应具备高并发处理能力、容错能力和灾备恢复能力，能够支撑多主体、大规模数据资产的整合与挖掘需求。

5.3.2 应建立健全的认证授权体系，对参与数据资产整合与挖掘的所有主体进行身份管理和权限控制，包括用户管理、角色管理、机构管理、认证管理、权限管理等，确保每个主体只能访问和操作其权限范围内的功能和数据。

5.3.3 应制定数据资产整合与挖掘的运营管理制度和流程，包括数据注册审核、协议审核、数据商店运营、计费计量、清结算、纠纷仲裁等，规范数据资产的流通和运营行为，保障参与主体的合法权益。

5.3.4 应建立监控审计体系，对数据资产整合与挖掘过程中的数据流转、操作行为、协议执行、系统性能等进行实时监控和审计。当发现异常事件时，应及时发出告警并启动应急响应流程。

5.3.5 应定期对空间云端平台和数据客户端进行技术升级和安全加固，修复已知漏洞，更新安全策略，确保平台和客户端的技术先进性和安全性。可为参与主体提供技术支持和培训服务，解答技术疑问，协助解决操作问题。

5.4 数据监管方

5.4.1 应具备对数据资产整合与挖掘过程的监管能力，能够接入空间云端的监控审计系统，获取数据资产的注册信息、流通日志、使用日志、协议执行情况等监管数据，实现对数据资产全生命周期监管。

5.4.2 应制定数据资产整合与挖掘的监管标准和规范，明确监管指标和监管流程，定期对参与主体的数据操作行为和平台运营情况进行检查和评估。

5.4.3 当发现数据资产整合与挖掘过程中存在违法违规行为时，应及时责令相关主体整改，并按照法律法规进行处理，必要时可暂停相关主体的数据操作权限，封存相关数据资产。

5.4.4 应建立监管信息共享机制，与数据提供方、消费方、运营方等主体保持沟通，及时反馈监管意见和建议，接受社会监督，处理相关投诉和举报。

5.5 数据开发者

5.5.1 应通过空间云端完成开发者资质申请，提供身份证明、技术能力证明、合规承诺。运营方审核通过后，应为开发者分配专用开发账号与客户端权限。

5.5.2 应在空间云端提供的数据沙箱或专用开发环境中开展开发工作，禁止将数据导出至沙箱外未授权环境。开发环境应支持隐私计算、洋葱加密、应用程序编程接口（API）编排等技术，具备日志记录功能（记录算法开发、模型训练、数据调用操作），日志同步至云端审计系统。

5.5.3 开发过程中使用的数据，应从数据商店申请获取，签署数据使用协议，明确开发用途，不得超范围使用数据。

6 数据架构

数据资产整合与挖掘的数据架构见图1。主要围绕数据空间参与者（提供方、消费方）与支撑模块（空间服务端、治理与监管）构建，形成“本地数据-共享数据-空间服务-消费使用”的全流程闭环。架

构以数据客户端、空间服务端为核心载体，通过多通道共享、治理管控等机制，实现数据在安全可控前提下的整合与挖掘利用，适配公共服务领域集中化管理与安全需求。

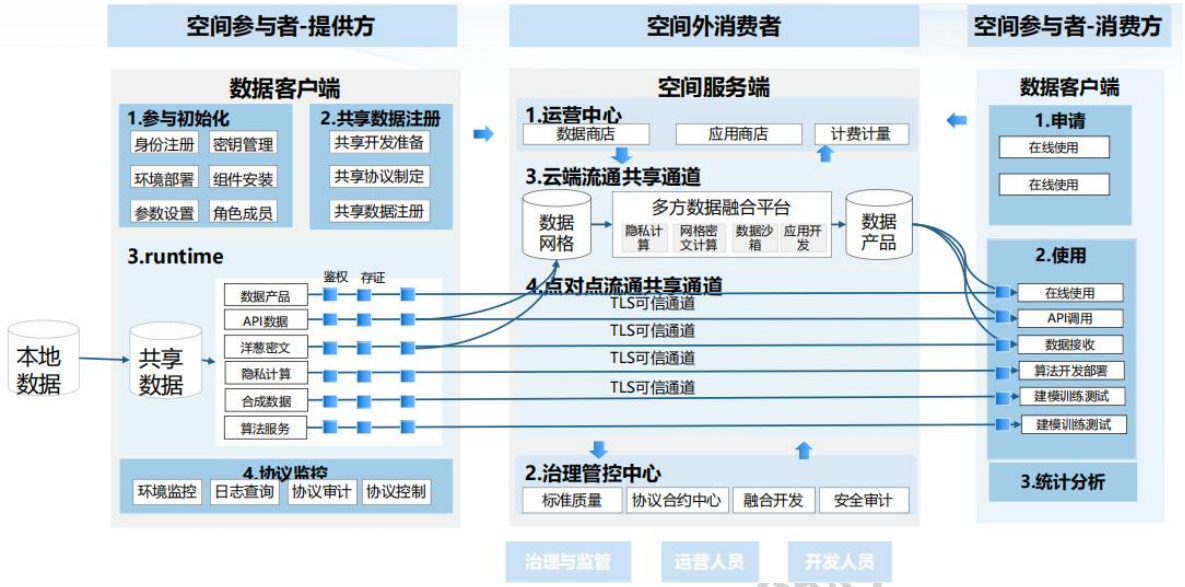

图 1 数据资产整合与挖掘数据架构

7 技术架构

7.1 总体架构

数据资产整合与挖掘的技术架构见图2。该架构主要采用“客户端-云端”协同架构，架构兼容中心化与去中心化两种部署模式：中心化模式下，数据资产可在云端进行统一托管（需采用加密存储），支持多主体集中访问与管理；去中心化模式下，数据资产主要存储在客户端侧，云端仅提供协议控制、身份认证、日志存证功能，数据传输通过点对点加密通道实现。

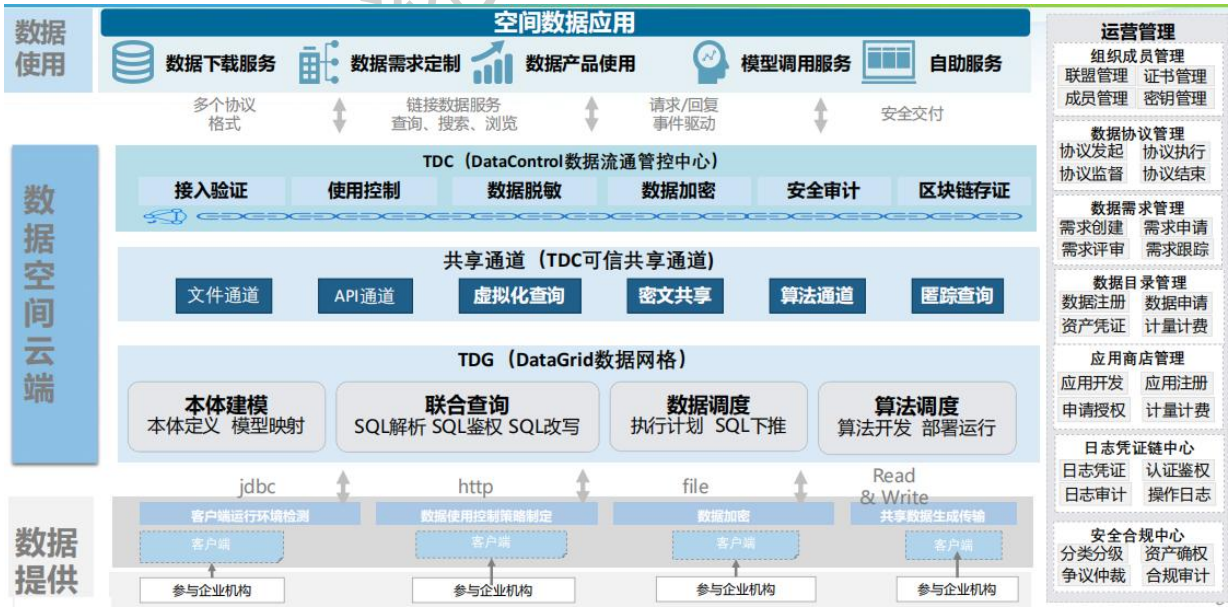

图 2 数据资产整合与挖掘技术架构

7.2 核心技术模块

7.2.1 数据加密

7.2.1.1 洋葱加密技术：对数据字段分层加密，支持不同使用场景配置差异化加密算法，实现“一数多模”安全使用。

注：洋葱加密技术主要采用确定性加密（DET）、保序加密（OPE）和同态加密（HOM）等多种加密算法组合应用。

示例：标识类字段（如用户 ID、企业统一社会信用代码等）采用 DET，保证相同明文始终加密为相同密文，支持跨数据表的关联操作；数值类字段（如交易金额、用电量等）采用 OPE，在密文状态下仍可进行大小比较；统计类字段则采用 Hom，支持密文状态下的加减乘除运算。

7.2.1.2 传输加密：采用安全套接层协议（SSL）/传输层安全协议（TLS）、虚拟专用网络（VPN）等技术保障数据传输过程安全，防止数据被截获或篡改；存储加密支持透明加密（TDE），确保密文数据仅通过授权客户端可访问。

7.2.2 隐私计算

7.2.2.1 联邦学习：支持多主体在数据不出域的情况下联合建模，通过参数交换、梯度聚合实现模型训练，适用于跨企业风控、用户画像等场景。

7.2.2.2 多方安全计算（MPC）：通过密码学算法实现多参与方协同计算，输出结果可见但原始数据不可见，适用于联合统计、匿踪查询等需求。

7.2.2.3 隐匿查询：通过加密查询条件、结果脱敏等技术，实现使用方在不暴露查询意图、提供方不暴露原始数据的前提下完成数据查询，支持结构化与非结构化数据查询。

7.2.3 协议控制

7.2.3.1 协议规则库：内置数据使用规则（使用方身份可信、数据有效期、使用次数、下载权限、脱敏方式等），支持自定义扩展，实现规则的技术化执行。

7.2.3.2 智能合约：基于区块链部署数据使用协议，自动触发协议条款。

7.2.3.3 实时监控：通过协议执行日志实时监测数据使用行为，当发现超协议操作时，应自动触发告警、冻结或销毁机制。

7.2.4 区块链存证

7.2.4.1 部署联盟链或私有链，存储数据资产凭证（数据权属、质量评估、授权信息）、操作日志（查询、计算、销毁记录）、合规证明（数据主体同意书、审计报告），确保数据不可篡改、可追溯。

7.2.4.2 支持与第三方审计平台对接，提供标准化存证接口，满足监管机构、行业协会合规检查需求。

7.2.5 数据网格

7.2.5.1 数据虚拟化引擎：实现异构数据源（关系型数据库、非关系型数据库、数据仓库、文件系统）的统一接入与映射，支持结构化查询语言（SQL）下推、查询重写优化，提升数据访问效率。

7.2.5.2 元数据管理工具：提供元数据采集（自动扫描、API 接入）、清洗、关联功能，支持数据溯源分析，帮助用户理解数据来源与流转路径。

7.2.5.3 数据调度中心：基于任务优先级、资源负载动态分配计算资源，支持定时调度与事件触发调度，保障数据时效性。

7.2.6 客户端

7.2.6.1　可信运行环境：采用安全容器技术，隔离数据处理环境，防止未授权访问与篡改，支持环境完整性检测。

7.2.6.2　本地计算引擎：集成 API 编排、数据融合计算（ETL+）、建模工具，支持密文数据本地分析，减少数据传输依赖。

7.2.6.3　日志审计组件：实时记录客户端数据操作（查询、下载、计算），生成标准化审计日志，同步至区块链存证模块，支撑合规追溯。

8　数据流转模式

8.1　密文流转

原始明文数据仅存储于提供方本地，通过洋葱加密、对称加密算法（AES）/公钥加密算法（RSA）等算法生成密文后传输至使用方或云端，计算过程不解密，保障数据隐私。

8.2　按需调度

基于数据使用协议（如使用次数、有效期、查询模式），由空间数据网格动态调度数据资源，避免无意义的数据拷贝与传输。

8.3　全链路存证

数据流转各环节（供给、接收、查询、计算、销毁）日志通过区块链存证，实现操作可追溯，支撑合规审计与争议仲裁。

9　数据源接入要求

9.1　接入范围

常见公共服务领域数据源包括以下类型，接入前应明确数据权属、敏感级别及使用范围，禁止接入无合法来源或超出授权范围的数据：

——政务服务数据：人口户籍、法人信息、政务审批记录、社保缴费数据、不动产登记数据等；

——医疗服务数据：患者诊疗记录、电子病历、医疗设备运行数据、公共卫生监测数据等；

——教育领域数据：教育资源分布、学生学籍信息、教学质量评估数据、校园安全管理数据等；

——城市规划数据：地理信息数据、城市基础设施分布数据、人口流动监测数据、土地利用规划数据等；

——第三方数据：企业信用数据、公共服务评价数据、地理信息数据等。

9.2　接入方式

9.2.1　实时性要求高的场景宜采用消息队列遥测传输（MQTT）协议流式接入，数据传输频率可根据场景调整。

9.2.2　批量数据宜采用定时文件导入，支持 CSV、Excel、JSON 格式。

9.2.3　跨部门实时互通数据宜采用 API 接口对接，API 接口协议符合 REST 规范，请求方法采用 GET（查询）/POST（提交），响应格式统一为 JSON 或 XML，接口调用通过身份认证。

9.2.4　第三方技术服务机构提供的数据宜通过标准化 API 接口或加密文件导入方式接入，禁止直接数据库直连。

9.3 接入流程

9.3.1 接入申请：数据提供方通过客户端提交接入申请，填写数据源类型、存储位置、数据范围、质量指标、使用协议等信息，上传合规证明材料。
9.3.2 审核验证：运营方对申请材料进行合规审核，技术团队通过测试环境验证数据源接入可行性（接口连通性、数据采集效率、质量达标情况），审核通过后生成接入配置方案。
9.3.3 配置部署：根据接入方案，在客户端或空间数据网格配置数据源连接参数（IP、端口、账号密码）、数据同步策略（增量/全量、同步频率）、元数据映射规则，完成数据源接入。
9.3.4 验收上线：接入完成后，进行数据采集测试（采集成功率、数据完整性）、功能测试（查询响应时间、计算准确性），验收通过后正式上线，纳入数据资产统一管理。

9.4 接入后管理

9.4.1 动态监控：应实时监控数据源连接状态、数据同步进度、数据质量变化。当出现连接中断、数据质量下降时，应自动触发告警并尝试重试或切换备用链路。
9.4.2 定期更新：根据业务需求变化，应支持数据源范围调整、同步策略优化、元数据更新，并通过审批流程（提供方申请→运营方审核→技术方执行）确保变更合规。
9.4.3 下线机制：当数据源不再使用或不符合合规要求时，应发起下线申请，完成数据备份、关联业务迁移后，删除接入配置，同时在区块链记录下线日志，确保全生命周期闭环。

10 数据资产整合流程

10.1 数据准备阶段

10.1.1 需求分析：明确数据整合目标、业务场景、数据范围，输出需求规格说明书。
10.1.2 方案设计：基于需求制定整合方案，包括数据架构选型（中心化/去中心化）、技术路线（密文整合/隐私计算整合）、资源配置（计算节点、存储容量、网络带宽）、进度计划，明确各参与方职责（提供方负责数据准备、使用方负责需求确认、运营方负责协调监督）。
10.1.3 环境搭建：部署客户端与空间数据网格相关组件，配置网络环境（防火墙、端口开放）、安全策略（访问控制、加密参数），搭建测试环境用于方案验证。

10.2 数据采集阶段

10.2.1 采集配置：根据数据源类型选择采集方式（批量采集、实时采集、增量采集），配置采集任务，设置数据过滤规则。
10.2.2 数据传输：采用加密传输技术将采集数据传输至客户端或空间数据网格，实时监控传输进度与成功率，确保数据无丢失、无篡改。
10.2.3 采集校验：对采集数据进行完整性校验、格式校验，校验不通过的数据进入异常队列，触发告警并通知相关方处理。

10.3 数据预处理阶段

10.3.1 数据清洗：通过规则引擎（如缺失值填充、异常值剔除、重复值去重）处理数据质量问题。
10.3.2 数据转换：根据整合需求进行数据格式转换、字段映射、数据脱敏，脱敏规则需符合数据使用协议。
10.3.3 数据融合：通过关联分析、数据聚合实现多源数据融合，生成标准化数据集，支持后续挖掘分析。

10.3.4　质量评估：对预处理后的数据进行质量评估，输出质量报告（包括完整性、准确性、一致性、时效性得分），评估达标后方可进入挖掘阶段；不达标数据应重新清洗或补充采集。

10.4　数据整合交付阶段

10.4.1　数据注册：将预处理后的数据集注册至空间数据网格，完善元数据信息（数据描述、字段说明、质量指标、使用协议），生成数据资产凭证，纳入数据商店管理。
10.4.2　授权配置：根据数据使用协议，配置数据访问权限、使用限制，通过智能合约固化授权规则。
10.4.3　交付验证：消费方通过客户端访问整合后的数据集，验证数据可用性、合规性，验证通过后完成交付，交付结果通过区块链存证。

11　挖掘算法选型

11.1　算法部署

11.1.1　本地部署：轻量级算法（如逻辑回归、K-Means）应部署于数据消费方客户端，基于密文数据本地计算，适用于数据量小、实时性要求高的场景。部署前应通过客户端环境检测，确保算法运行稳定性；计算过程中应实时记录资源占用，避免影响客户端其他业务。
11.1.2　云端协同部署：复杂算法（如深度学习、联邦学习）应采用“客户端本地计算+云端协同调度”模式。部署应配置安全通信通道，确保参数传输过程不可篡改；云端应支持多节点负载均衡，避免单点故障。
11.1.3　沙箱部署：高敏感数据挖掘应在安全沙箱内部署算法，沙箱具备环境隔离（与外部网络物理隔离）、操作审计（实时记录算法调用、数据访问）、数据防泄漏（禁止未授权下载、拷贝）功能，确保挖掘过程全程可控。

11.2　算法成果管理

11.2.1　模型注册：挖掘生成的模型应注册至空间数据网格的模型商店，完善元数据信息（模型名称、算法类型、输入输出格式、训练数据来源、性能指标），生成模型资产凭证，支持模型版本管理（版本迭代、历史版本回溯）。
11.2.2　模型发布：模型发布前应通过安全审核，审核通过后可通过 API 接口或软件开发工具包（SDK）形式开放给使用方调用；发布时应配置使用权限，通过智能合约控制模型使用行为。
11.2.3　模型监控与更新：应实时监控模型调用情况（调用频率、响应时间、错误率）与效果衰减情况。当模型效果低于阈值时应触发更新提醒，支持基于新数据重新训练模型并迭代发布，确保模型时效性与准确性。

12　安全管控要求

12.1　身份认证与权限管控

12.1.1　多层级身份认证

12.1.1.1　可采用“用户名密码+动态令牌+数字证书”三重认证机制，确保用户身份真实性；针对高权限操作（如数据销毁、协议修改），应额外进行实人认证（人脸识别、身份证核验）。
12.1.1.2　应支持与企业统一身份认证系统对接，实现跨系统身份同步，避免重复注册；身份信息应加密存储，禁止明文存储敏感信息。

12.1.2 权限管控

12.1.2.1 应基于“最小权限原则”，按角色（数据提供方、消费方、运营方、监管方）与职责分配权限。

示例：数据提供方可管理自有数据的授权与监控，消费方可调用授权数据与模型，运营方可进行审核与运维，监管方可查看审计日志。

12.1.2.2 应支持字段级、操作级权限控制。

12.1.2.3 权限变更应通过审批流程（申请人提交→审批人审核→系统执行），变更记录同步至区块链存证，确保权限操作可追溯、不可抵赖。

12.2 数据安全管控

12.2.1 数据存储安全

12.2.1.1 原始明文数据应仅存储于数据提供方本地或授权的加密存储设备，禁止存储于公共云或未授权服务器；密文数据存储应采用透明加密技术（TDE），密钥由提供方或第三方密钥管理系统（KMS）统一管理，定期更换密钥。

12.2.1.2 应支持数据备份与恢复机制，备份数据与原始数据物理隔离，备份过程加密传输与存储；应定期进行恢复测试，确保备份数据可用性，应对数据丢失风险。

12.2.2 数据传输安全

12.2.2.1 数据传输应采用 SSL/TLS 1.2 及以上版本加密协议，确保传输链路安全；针对大文件传输，应支持分段加密传输与校验，避免数据传输过程中丢失或篡改。

12.2.2.2 禁止通过未授权通道传输数据，所有数据传输应通过空间数据网格的可信通道进行，通道具备接入验证与异常检测功能，拦截非法传输请求。

12.2.3 数据使用安全

12.2.3.1 基于数据使用协议，应通过技术手段强制执行使用规则。

示例：数据有效期到期后自动删除，使用次数超限后禁止访问，未授权查询模式触发拦截。

12.2.3.2 应支持数据使用行为监控，实时检测异常操作（如高频次查询、跨地域访问、非工作时间访问）。当发现异常时应触发告警（短信、邮件）并暂停操作权限，待人工审核通过后恢复；异常行为应记录纳入审计日志，支撑后续风险分析。

12.2.3.3 禁止数据再流通，应通过客户端技术限制密文数据的拷贝、转发与导出。

12.3 审计与合规管控

12.3.1 全链路审计日志

12.3.1.1 应记录数据资产整合与挖掘全流程操作日志，包括：数据源接入日志（接入时间、接入人、数据范围）、数据操作日志（查询、下载、计算、销毁的时间、主体、内容）、权限变更日志（权限分配、变更、回收的审批人与执行结果）、模型调用日志（调用主体、调用次数、响应结果）。

12.3.1.2 审计日志应标准化格式，包含唯一标识、操作主体、操作时间、操作内容（数据 ID、模型 ID、权限范围）、操作结果（成功/失败、错误码），确保日志完整性与可分析性。

12.3.1.3 审计日志应实时同步至区块链存证，不可篡改、不可删除；应支持日志查询与导出功能，满足监管机构的合规检查需求。查询时应进行身份认证与权限校验，禁止未授权访问。

12.3.2 合规检测与风险预警

12.3.2.1　应内置合规检测规则库，定期对数据资产整合与挖掘过程进行合规扫描。

12.3.2.2　合规检测重点包括：数据来源合规性（是否取得数据主体同意、是否具备权属证明）、数据使用合规性（是否符合授权范围、是否超协议使用）、数据流转合规性（是否符合跨境数据传输要求、是否通过可信通道传输）。

12.3.2.3　当检测到合规风险时，应自动触发风险预警，分级推送至相关责任方（数据提供方、消费方、运营方），并提供整改建议；应跟踪整改进度，确保风险闭环处理。