ICS xx
CCS xx

团 体 标 准

T/ACEF XXXX—2024

生态环境大数据管理平台架构技术规范

Technical specification for architecture of ecological environment big data management platform

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

2024-XX-XX 发布 2024-XX-XX 实施

中 华 环 保 联 合 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。

本文件由北京环丁环保大数据研究院提出。

本文件由中华环保联合会归口。

本文件起草单位：北京环丁环保大数据研究院、北京市生态环境监测中心、北京微芯区块链与边缘计算研究院、联通数字科技有限公司。

本文件主要起草人：…、…、…。

生态环境大数据管理平台架构技术规范

1 范围

本文件规定了生态环境大数据管理平台架构设计的原则及架构各组成部分和各部分的功能与实现要求。

本文件适用于环保相关部门开展生态环境大数据管理平台的设计、建设、管理、应用和运维的指导。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

HJ/T 418　　环境信息系统集成技术规范

HJ/T 720　　环境信息元数据规范

《生态环境大数据建设总体方案》（环办厅[2016] 23号 ）

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据质量 data quaility

在指定条件下使用时，数据的准确性、完整性、一致性、可信度、及时性和可用性等方面满足明确的和隐含的要求的程度。

3.2

软件开发工具包 Software Development Kit, SDK

辅助实现软件产品某些功能的相关文档、范例和工具的集合

3.3

应用程序接口 Application Programming Interface, API

一组定义了软件组件之间交互的规范和方法的集合。

4 总体原则

4.1 模块化与服务化原则

倡导将复杂系统分解为若干独立且协同的模块，每个模块聚焦于特定功能，促进系统的灵活构建与高效维护。鼓励采用微服务架构，将平台功能以服务形式封装，促进服务的独立演进与快速响应市场变化。

4.2 可扩展性与灵活性原则

设计时考虑未来数据增长和业务扩展需求，确保平台能够灵活应对规模变化，支持水平或垂直扩展。构建灵活可调的架构，以便轻松集成新技术、适应新需求，确保系统的持续进化能力。

4.3 安全与稳定原则

构建全方位的安全防护体系，包括数据加密、访问控制、审计追踪等，保障数据全生命周期的安全。确保平台在各类复杂场景下稳定运行，通过冗余设计、负载均衡等策略提升系统可用性。

4.4 开放共享与规范化原则

倡导数据资源的开放共享，推动跨领域、跨组织的数据合作，促进数据价值的最大化利用。遵循并推动数据标准制定，促进数据格式的统一与互操作性，提升数据质量与可信度。

4.5 前瞻性与创新性原则

保持对技术发展趋势的敏锐洞察，确保平台架构能够引领而非追随技术发展潮流。鼓励技术创新与应用探索，将新技术融入平台建设中，不断提升平台的智能化水平和用户体验。

5 总体架构

依据《生态环境大数据建设总体方案》（环办厅[2016] 23号 ），在生态环境大数据总体架构中，包括生态环境大数据管理机制、生态环境大数据标准规范体系、运维和信息安全体系以及大数据环保云平台、大数据管理平台、大数据应用平台等内容。生态环境大数据管理平台是生态环境大数据总体架构的重要组成部分，属于生态环境大数据的数据资源层，为云平台的基础上为应用平台提供数据采集、预处理、存储、分析等支撑服务。生态环境大数据管理平台总体架构的设计应按图 1实施。

图 1 大数据管理平台总体架

6 大数据标准规范

6.1 标准体系建设

应根据大数据管理平台总体架构、建设内容、政策法规要求梳理标准化需求，制定所需的标准体系，配置与体系对应的标准文献。

6.2 标准更新

应根据法规、技术进步、标准更新及时同步更新标准体系：

a) 标准体系框架更新；
b) 标准明细表更新；
c) 配置更新标准文献

6.3 标准持续改进

应按PDCA组织标准体系实施计划、实施、检查、改进。

7 数据接入

按《生态环境大数据平台数据接入规范》的相关要求实施。

8 数据抽取、转换与加载

8.1 数据抽取、转换与加载要求

数据抽取、转换与加载的要求如下：

a) 应确定数据提取的周期、频率和方式及增量提取或全量提取设定；
b) 应确定数据转换的规则和处理逻辑，包括数据清洗、数据修复、数据映射、数据合并等；
c) 应确定数据加载的目标系统和方式，包括批量加载和实时加载；
d) 应确定数据质量的标准和监控方法，包括数据完整性、准确性和一致性等要求；
e) 应确定 ETL 过程的性能指标，包括数据处理的速度、吞吐量和响应时间等。

8.2 数据抽取、转换与加载步骤

8.2.1 数据抽取

数据抽取步骤如下：

a) 识别和验证数据源，确保数据源的可用性和准确性，明确数据源的架构、模式以及数据的关系和依赖；
b) 通过适当的工具和技术建立与数据源的连接，从数据源中提取所需的数据；
c) 数据进行校验、清洗及验证，保证数据的完整性、准确性和一致性等要求，并清除无效或重复数据。

8.2.2 数据转换

数据转换步骤如下：

a) 根据数据需求和业务规则对数据进行筛选和过滤，只提取和转换与业务目标相关的数据；
b) 将来自不同数据源的数据进行映射与整合，将多个表或文件中的相关数据进行匹配和关联，根据业务需求对数据进行合并；
c) 数据转换和计算，包括格式转换、单位转换、数据标准化、数据计算等操作；
d) 数据清洗和规范化，修复数据中的错误、缺失值和不一致性，并将数据转换为目标系统所需的标准格式和结构；
e) 从其他数据源或外部系统获取额外的数据对目标系统进行数据补充和扩展。

8.2.3 数据加载

数据加载步骤如下：

a) 创建目标表、定义表的结构和关系，确保目标系统具备足够的容量和性能来处理加载的数据；
b) 将数据逐行或批量加载到目标表中；
c) 数据校验和验证以确保数据的准确性和完整性等质量要求，比较源数据和目标数据，检查数据的一致性；
d) 建立错误处理机制，解决加载错误和故障，并将错误信息记录到日志中。

8.3 数据抽取、转换与加载维护

8.3.1 范围

数据抽取、转换与加载维护包括以下三部分内容

8.3.2 日志检查

日志文件报告。监测数据ETL过程日志，形成汇总报告和详细报告。汇总报告应包括作业总数、警告作业数、成功作业数、失败作业明细、开始时间、结束时间等内容。详细报告应包括作业名称、开始时间、结束时间、运行状态等内容。

8.3.3 出错处理

数据ETL过程错误分为分为以下三类：

a) 抽取错误：简称 E（Extract）类错误
b) 转换错误：简称 T（Transformer）类错误
c) 装载错误：简称 L（Loading）类错误

应根据ETL实际故障原因编制相应的错误处理脚本。

8.3.4 备份与恢复

ETL备份应包括运行环境备份及数据备份，运行环境备份应保证两个或两个以上相对独立且具有配置相同的ETL设备互为备份；数据备份应每天保留一个备份文件，至少保留7天。

ETL恢复应包括运行恢复和数据恢复。

9 数据存储存证

9.1 数据存储

9.1.1 数据存储要求

数据存储要求如下：

a) 支持高物理安全性、高容量的数据存储以及高存取性能等；
b) 支持数据加密、数据审计、灾难恢复等技术保障数据安全，支持多级用户权限管理；
c) 支持多维结构的数据模型，通过分层存储、元数据管理、维度管理等技术，实现高效的数据分析及查询；
d) 满足联盟链数据存储的要求。

9.1.2 数据存储架构

生态环境大数据平台大数据存储子系统提供大数据的分布式存储管理，涵盖多种存储方式，包括分布式文件存储、分布式结构化数据存储、分布式列式数据存储、分布式图数据存储以及区块链数据存储，总体框架见图 2。

图 2 大数据存储子系统总体框架

生态环境大数据平台采用分布式松耦合联邦集群方式进行数据存储，数据存储技术架构如图 3所示。

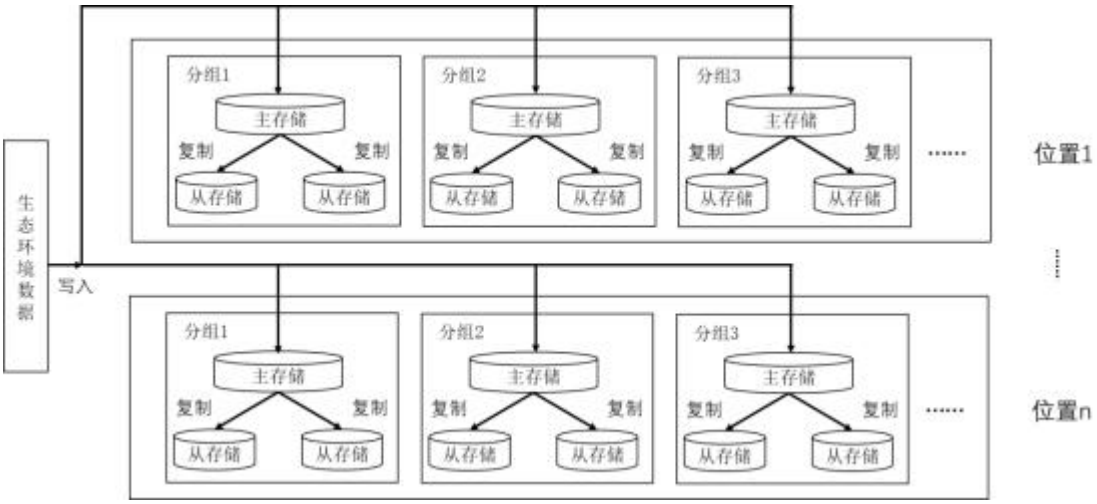

图 3 大数据平台数据存储技术架构

9.2 数据存证

9.2.1 数据存证要求

数据存证要求如下：

a) 数据的原文或完整性校验值、附属信息等数据同步存储至区块链存证平台；
b) 对于原文存证的数据可保证数据的完整性，对于校验值存储的数据可验证数据的完整性。

9.2.2 数据上链存储

数据上链存储包括以下步骤：

a) 数据哈希生成：使用哈希算法生成数据的唯一哈希值，哈希值作为数据的“指纹”；
b) 交易创建：将数据的哈希值作为交易的一部分，创建区块链交易；
c) 共识机制：通过区块链网络的共识机制验证交易；
d) 时间戳记录：每笔交易都有一个时间戳，记录数据存证的时间点；
e) 存储上链：验证并加盖时间戳的数据存储在区块链网络中，所有节点都持有数据副本，确保数据的高可用性和可靠性。在实践中，因区块链存储负担较重，可以将体量较大的实际数据存储在链下，链上只存储实际数据的哈希值和相关元数据，以减少区块链的存储负担。

9.2.3 数据验证

数据验证包括以下内容：

a) 哈希变换：验证数据时，将待验证数据重新进行哈希运算，获取哈希值；
b) 一致性检查：将待验证数据的哈希值与区块链上保存的哈希值进行比对，若一致，则证明数据未被篡改，反之则说明数据被篡改。

10 数据治理

10.1 元数据管理

10.1.1 元数据管理要求

遵循HJ 720相关要求。

10.1.2 元数据管理内容

元数据应描述数据的格式、结构、值域、来源、血缘沿袭、定义及预期用途等内容。元数据可分为技术元数据、业务元数据和管理元数据三类来管理。元数据管理包括三部分内容：

a) 元数据标准化：应符合 HJ 720 的相关要求来描述和格式化元数据；
b) 元数据基本管理：包括元数据的增加、删除、修改、查询及统计和使用情况分析等功能，元数据的增加应符合 HJ 720 的相关要求；
c) 元数据分析：环境元数据分析包括血缘分析、实体关联分析、实体影响分析、指标一致性分析。

10.2 主数据管理

10.2.1 主数据管理要求

主数据管理要求如下：

a) 规范化、集中化、维护和更新；
b) 保证主数据的一致性、准确性和可靠性。
c) 主数据必须符合组织制定的数据命名、命名规则和格式要求；
d) 对主数据的变更和操作，应有记录和审计跟踪。

10.2.2 主数据管理内容

主数据管理是一种管理组织中关键业务数据的综合方法，包括以下内容：

a) 主数据定义：按照预定义的主数据标准确定数据的结构，包括实体、属性、关系和约束；
b) 主数据整合与集成：从多个数据源（如内部系统、外部合作伙伴、第三方数据提供商）收集主数据并将来自不同源的数据整合在一起，解决数据冗余和冲突；
c) 主数据质量管理：采用数据质量评价指标和评价方法持续监控主数据质量；
d) 主数据生命周期管理：管理数据的整个生命周期，从创建、使用、共享到归档和删除。

10.3 数据生命周期管理

通过定义数据生命周期阶段、制订数据管理政策、确定数据保护措施、建立数据治理框架的步骤编制数据生命周期管理策略，实施数据生命周期管理。

10.4 数据质量管理

10.4.1 数据质量评价维度

数据质量评价维度应符合DB15/T 1873-2020 第5章的要求。

10.4.2 数据质量控制内容

数据质量控制内容应符合DB15/T 1873-2020，6.2-6.4的要求。

10.4.3 数据质量评分方法

数据质量评分方法应符合DB15/T 1873-2020第7章的要求。

11 数据服务

11.1 数据服务技术要求

数据服务的技术要求如下：

a) 本标准中的数据交换格式遵循 HJ/T 418-2007，6.2.3.1 规定和要求；
b) 本标准中的数据交换技术遵循 HJ/T 418-2007，6.2.3.2 规定和要求；
c) 应具有数据访问权限验证、访问控制、服务记录、溯源和目录链数据共享等功能；
d) 各类数据服务应屏蔽不同的中间存储，提供是格式规范化的统一 API，构建 API 集市，数据服务用户可以查看 API 集市现有 API，通过申请相关权限使用 API，也可根据实际需要编制新 API 经注册登记后纳入 API 集市；
e) 应提供提供智能合约 SDK 及 API 服务供数据服务用户实现区块链数据的读取、写入与验证。

11.2 数据服务内容

数据服务内容如下：

a) 为空气、水、土壤、自然生态、核与辐射、污染源、行政办公等数据及公众数据或文件提供统一的认证和接口；
b) 数据订阅；
c) 目录链数据共享；
d) 数据服务记录及数据溯源查询；
e) 数据服务用户定制服务。

12 数据安全

12.1 数据安全要求

大数据管理平台数据安全应满足以下要求：

a) 支持合规的数据访问并防止不合规的访问；
b) 支持对隐私、防篡改、保密等制度、法规的遵循；
c) 确保满足利益相关方对隐私、保密及数据权限的要求。

12.2 数据管理安全体系

本文件规定了大数据管理平台的平台安全和数据安全，不涉及基础设施安全和应用安全。应包括但不限于认证安全、鉴权安全、传输安全、分类分级、存储安全、交换共享安全、管理安全、审计安全、备份与恢复等数据管理安全体系如图 4所示。

图 4 数据安全管理体系

13 平台运维

13.1 运行维护要求

为保障生态环境大数据管理平台的长期稳定运行，应制定详细的维护方案，建立完善的维护体系。平台运维的要求：

a) 服务、管理并重。
b) 完善的数据安全、账号安全、资源保障、应急处理策略。
c) 建立规范化的管理运维问题处理流程、需求审批流程。

d) 明确责任和奖惩。

13.2 运行维护体系

生态环境大数据平台的运行维护体系包括运行维护制度、运行维护流程、运行维护团队、运行维护方案、运行安全保障五大部分。运行维护体系如图 5 所示。

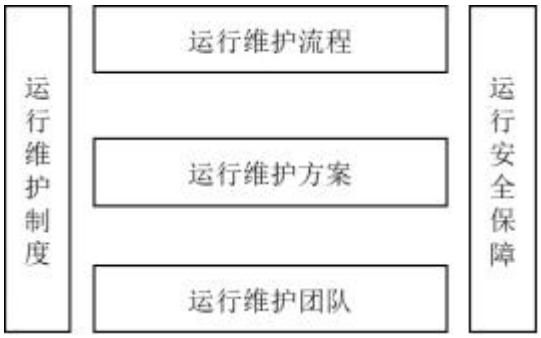

图 5 大数据管理平台运行维护体系

13.3 运行维护制度

为保证项目运维正常开展，应制定以下运行维护制度：

a) 系统备份和恢复制度：包括系统备份的策略及执行规定、备份体系的搭建和描述、日常记录及恢复测试；
b) 账号安全管理规定：包括账号审批原则、账号使用约束规定；
c) 需求处理流程制度：包括需求处理原则、需求提出、审批，问题处理流程，应急预案；
d) 问题处理制度：包括问题解决流程描述、问题解决责任界定、问题归档和总结提炼。

参考文献

[1] GB/T 22080 - 2008 信息技术 安全技术 信息安全管理体系[S].
[2] GB/T 37973 - 2019 信息安全技术 大数据安全管理指南[S].
[3] GB/T 38633 - 2020 信息技术 大数据 系统运维和管理功能要求[S].
[4] GB/T 38666 - 2020 信息技术 大数据 工业应用参考架构[S].
[5] HJ 718 - 2014 环境信息共享互联互通平台总体框架技术规范[S].
[6] HJ 729 - 2014 环境信息系统安全技术规范[S].
[7] HJ 417 - 2007 环境信息分类与代码[S].
[8] HJ/T 418 - 2007 环境信息系统集成技术规范[S].
[9] HJ/T 720 - 2014 环境信息元数据规范[S].
[10] DB15/T 1873 - 2020 大数据平台 数据接入质量规范[S].
[11] 环办厅[2016]23号 生态环境大数据建设总体方案[S].