ICS 点击此处添加 ICS 号
CCS 点击此处添加 CCS 号

NY

中 华 人 民 共 和 国 农 业 行 业 标 准

NY/T XXXXX—XXXX

农业农村大数据存储系统功能及性能要求

Features and performance requirements of agricultural and rural big data storage system

(点击此处添加与国际标准一致性程度的标识)

（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

中华人民共和国农业农村部　　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。

本文件由农业农村部大数据发展中心提出。

本文件由农业农村部大数据标准化委员会归口。

本文件起草单位：

本文件主要起草人：

引　言

由于不同行业数据具有领域特色，尤其是农业领域数据多源异构性特点明显，现有通用性标准在指导农业农村领域大数据存储应用方面尚存在一些不足。一是农业农村数据来源多样、数据类型众多，包括调查统计数据、 业务系统数据以及遥感、传感器、智能终端等技术装备自动产生数据，各类数据时空维度、体量、更新周期的不同，对储存系统功能及性能要求不尽相同，通用标准因为领域针对性不足，标准缺乏可操作性。二是系统性能是衡量一个系统重要指标，存储系统功能与系统性能密不可分，二者应该统筹考虑。因此，有必要针对农业农村大数据特点，编制《农业农村大数据存储系统功能及性能要求》行业标准，指导不同层级、不同种类农业农村大数据中心建设，提高数据存储系统效率和质量。

本文共分为7个部分，其中范围、规范性引用文件、缩略语等对本文编制做了规范性基础要求，其次，设计了整体农业农村大数据存储系统架构，最后从农业农村大数据存储系统功能和性能方面要求进行了约定。

农业农村大数据存储系统功能及性能要求

1 范围

本文件规定了农业农村大数据存储系统的术语和定义、缩略语、农业农村大数据存储系统架构、农业农村大数据存储功能要求、农业农村大数据存储性能要求。

本文件适用于农业农村大数据存储系统的设计、开发、系统对接、测试和应用部署。

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。

GB/T 20988-2007 信息安全技术 信息系统灾难恢复规范

GB/T 25000.23-2019 系统与软件工程 系统与软件质量要求和评价（SQuaRE） 第23部分：系统与软件产品质量测量

GB/T 35295 信息技术 大数据 术语

GB/T 37732 信息技术 云计算 云存储系统服务接口功能

GB/T 37737 信息技术 云计算 分布式块存储系统总体技术要求

GB/T 37973 信息安全技术 大数据安全管理指南

GB/T 38676 信息技术 大数据 存储与处理系统功能测试要求

NY/T 4261 农业大数据安全管理指南

3 术语和定义

GB/T 35295和NY/T 4261界定的以及下列术语和定义适用于本文件。

3.1

大数据 big data

具有数量巨大、种类多样、流动速度快、特征多变等特性，并且难以用传统数据体系结构和数据处理技术进行有效组织、存储、计算、分析和管理的数据集。

[来源：GB/T 37973-2019,3.1]

3.2

农业农村大数据 agriculture and rural big data

指在涉农生产、经营、管理、服务过程中，制作或获取并以电子化或其他形式记录、保存的涉及农业农村资源、主体、产品且同时具有大数据特征的数据集。

3.3

分布式存储 distributed storage

一种将数据分散存储在多个物理位置的计算节点上，以提高数据可靠性、可用性和处理效率的技术架构。

4 缩略语

下列缩略语适用于本文件。

API：应用程序接口（Application Programming Interface）

DAS：直连存储（Direct-Attached Storage）

NAS：网络附加存储（Network Attached Storage）

SAN：存储区域网络（Storage Area Network）

5 农业农村大数据存储系统架构

农业农村大数据存储系统架构主要考虑农业大数据领域的存储和应用的有效对接，以及存储与应用系统之间的层级调用和资源管理方法，确保数据在存储环节的安全性和稳定性，同时保障其在应用过程中的高效性和准确性。整体架构见图1，从下到上可分为如下4层：

（1）存储资源层：农业农村大数据的物理存储资源层。可包括直连式存储、存储区域网络、网络附加存储、脱机存储媒体等。按照不同的数据划分方式选择存储资源，如按照数据访问频率、数据重要性等。

（2）存储系统层：农业农村大数据的分布式存储系统层。提供适用于结构化、非结构化及半结构化数据类型的存储系统。包括适用于农业农村大数据存储的各类存储系统，如关系型数据存储、列式数据存储、空间数据存储、图数据存储、时序数据存储、向量数据存储、文件数据存储、对象数据存储等。

（3）接口层：为应用层提供使用存储系统层服务的接口。主要是API、运维、系统服务、权限与安全4类。

（4）应用层：通过具体的应用程序满足农业农村领域的业务需求。

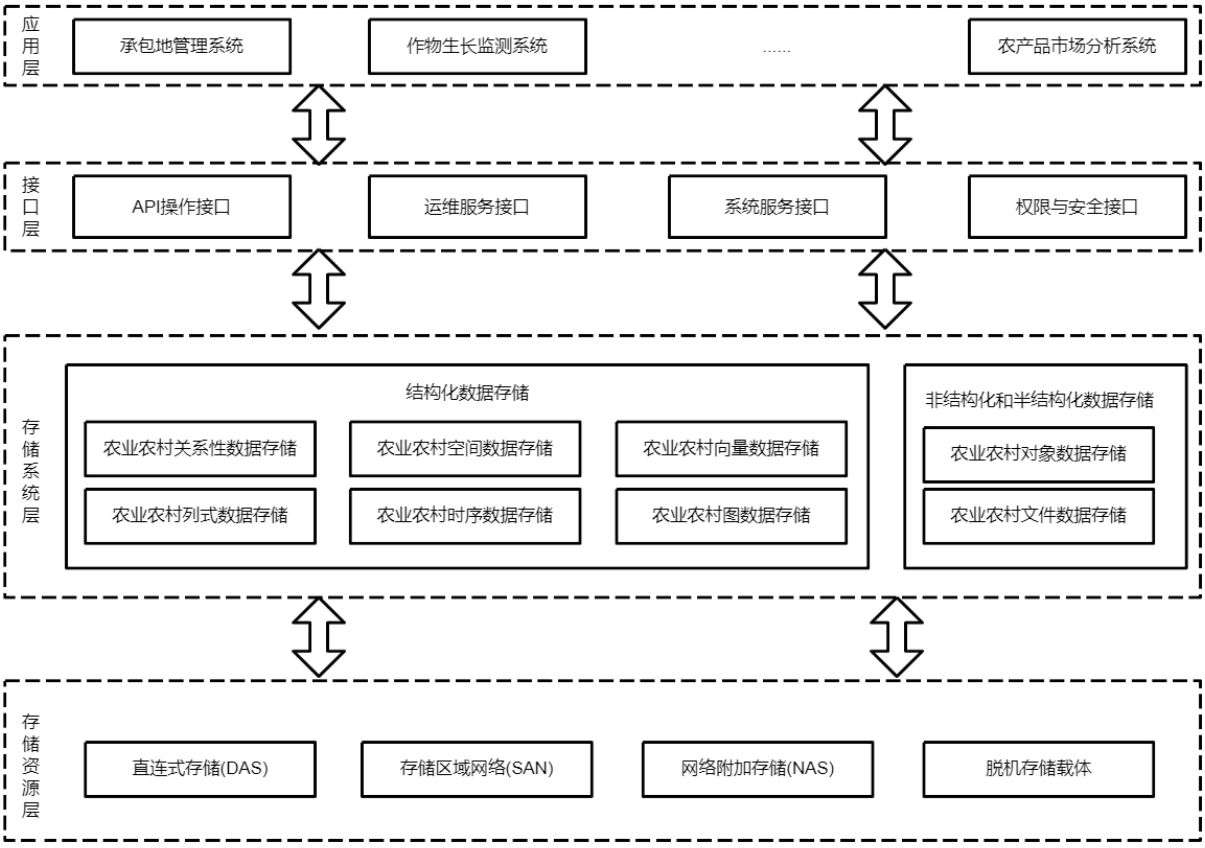

图 1 农业农村大数据存储系统架构图

6 农业农村大数据存储功能要求

6.1 一般要求

6.1.1 存储资源要求

a) 应具备使用不同存储设备作为物理存储介质的能力；
b) 应具备海量农业数据的存储和查询能力；
c) 应具备横向和纵向扩展能力，以适应不断增长的数据量和访问压力；
d) 应具备不同存储介质间数据平滑迁移、复制、转换的能力；
e) 应具备数据按冷热、分类分级存储后，数据仍可按统一出入口访问的能力；
f) 应具备完善的资源监控和管理功能，包括性能监控、故障报警、故障诊断、资源使用情况等；
g) 宜具备高效的自动化管理和弹性伸缩能力，且扩缩容操作不影响系统的可用性、可靠性及数据的完整性。

6.1.2 数据一致性和可靠性要求

a) 应保证数据的一致性，确保多个副本之间的数据准确无误；
b) 应提供可靠的数据存储，具备数据备份和恢复机制；
c) 宜具备数据多版本控制、快照、软删除等功能；
d) 宜具备数据多机房、多地备份能力；

e） 宜具备数据按分类分级选择存储策略的能力。

6.1.3 兼容性和开放性要求

a） 应能够与不同的硬件、软件和系统进行兼容和集成；
b） 应支持开放的标准和接口，便于与其他系统进行交互和数据共享。

6.1.4 数据安全性要求

a） 应具备细粒度的访问权限控制和多种方式的用户认证能力；
b） 应具备数据存储和传输的加密解密功能；
c） 应具备流量访问监控预警、审计日志、脱密脱敏等能力；
d） 应具备多租户管理和数据隔离能力；
e） 宜具备数据全生命周期的安全管理能力；
f） 宜具备使用校验和哈希函数等技术验证数据在传输和存储过程中是否被篡改。

6.1.5 管理和维护要求

a） 应提供直观的管理界面和工具，方便进行配置、监控和故障排查；
b） 应具备详细的设计文档、用户手册等；
c） 宜具备自动化运维、升级、补丁管理等能力。

6.1.6 能效和成本要求

a） 宜具备物理存储资源的绿色认证和优化的散热管理；
b） 宜具备可视化的效能监控面板；
c） 宜在满足性能和功能需求的前提下控制存储成本。

6.2 不同数据类型存储要求

6.2.1 文件存储功能

实现存储农业政策文件、农产品检测报告等以各种文件格式存储的数据。

a） 应具备文件的上传、下载、读写、复制、移动、删除、重命名等基本功能，并支持批量操作；
b） 应具备支持大容量文件的存储，能够满足长期的数据积累需求；
c） 应提供高效的文件检索功能，能够根据文件名、关键词、日期等属性快速查找文件；
d） 应支持文件分类和目录管理，方便用户组织和管理文件；
e） 宜支持文件跨平台访问，能够在不同的操作系统和设备上进行文件操作。

6.2.2 对象存储功能

实现数据、元数据和唯一标识符数据对象的存储，存储遥感影像、电子合同附件、农业监控设备产生的视频和图像数据等大量半结构化和非结构化数据。

a） 应具备存储桶的创建、删除、获取对象列表、查询具体信息、获取位置等操作；
b） 应具备对象的上传、下载、读写、复制、移动、删除、重命名、预请求、分块上传、分块下载、断点续传等操作；
c） 宜支持跨区域复制功能，包括开启、停止、复制的规则配置等功能；
d） 宜支持跨域资源共享功能。

6.2.3 关系型数据存储功能

基于关系模型对农业资源管理、农产品质量追溯等公共业务系统的数据进行组织存储。

a) 应支持创建、修改和删除表结构，包括定义列的数据类型（如整数、字符串、日期等）、约束条件（主键、唯一键、外键等）和默认值；
b) 应具备数据插入、更新、删除操作和基于标准结构化查询语言的查询语法；
c) 应支持事务、索引、触发器、视图、存储过程和函数等功能；
d) 宜支持读写分离、存算分离架构和支持分库分表。

6.2.4 空间数据存储功能

实现存储、检索和管理具有空间维度的地理信息系统数据、遥感影像等空间数据。

a) 应能够存储点、线、面等基本几何对象，支持复杂的空间数据结构，如拓扑数据结构、网络数据结构等；
b) 应具备准确处理和存储不同的地理坐标系和投影坐标系空间数据，支持坐标系的转换和动态投影；
c) 应具备高效的数据空间索引机制，建立有效的空间索引，如 R 树、四叉树等；
d) 应支持将空间对象的属性数据（如名称、类型、描述等）与几何数据紧密关联存储；
e) 宜支持空间信息数据可视化功能。

6.2.5 图数据存储功能

优化存储和查询图形结构数据，实现农业供应链分析、农业知识图谱等图结构数据的存储与处理。

a) 应具备灵活的图数据模型，能够表示节点、边及它们之间的复杂关系和属性；
b) 应支持以邻接表和邻接矩阵的形式存储图；
c) 应支持为节点和边存储丰富的属性数据，并能高效地进行读写操作；
d) 应支持高效地处理节点和边的添加、删除和修改操作，实时更新图的结构和属性；
e) 应具备高效的图遍历算法，以便快速查询图中的路径、节点和关系；
f) 应具备高效的索引机制，支持图的高效、准确查询，包括属性索引、图结构索引等；
g) 宜支持复杂图处理任务的能力，如图嵌入、图类聚、图分类等。

6.2.6 列式数据存储功能

实现农业生态环境数据、农村社会经济数据等分析型工作负载优化并按列进行存储。

a) 应具备按照列组织数据的能力，对于指定列的数据可进行高效读取和处理；
b) 应具备对列数据进行排序存储的能力；
c) 应支持对列数据的更新操作，同时保持数据的一致性和完整性；
d) 应支持面向列的高效压缩和数据聚合；
e) 宜具备存算分离架构。

6.2.7 时序数据存储功能

以时间戳为索引，高效存储和管理随时间变化的数据点，实现农业监测传感器数据、气象数据等随时间变化而进行变化的时间序列数据存储。

a) 应具备基于时间戳的数据索引功能；
b) 应具备根据时间维度或者其他维度的数据分片能力；
c) 应支持数据聚合和摘要，如计算时间窗口内的最大值、最小值、平均值等摘要信息；
d) 应支持根据时间维度和其他维度进行数据查询和过滤；
e) 应支持数据保留策略；

f） 宜具备缺失时间数据处理功能，如中位数填充，线性拟合填充等。

6.2.8 向量数据存储功能

通过分布式架构优化大规模向量数据的存储密度和查询速度，支持复杂的空间数据分析和机器学习任务，实现农业自然灾害预测预警、农产品产量预测等向量数据的存储和处理。

a） 应具备基于相似性的搜索功能，如余弦相似性、欧几里得距离等度量；

b） 应具备高效的向量索引结构；

c） 应支持与分布式计算框架集成。

7 农业农村大数据存储性能要求

7.1 存储系统性能指标

衡量存储系统性能主要可以从系统的可用性和维护性等方面考虑。详细性能测试指标及方法见表 1。

a） 应提供足够的存储空间以容纳不断增长的数据量，数据存储量可达TB级，同时支持灵活的存储空间分配和管理；

b） 应具备良好的I/O性能，包括数据读写速度、延迟、吞吐量等；

c） 延迟时间应尽量保持在较低水平，能够快速获取到所需数据结果；

d） 存储系统层应支持在主流的操作系统上安装、运行。

表1 性能测试指标及方法

一级指标	二级指标	描述	方法	参考数值
可用性	平均吞吐量	单位时间内完成作业的平均数量	使用GB/T 25000.23-2019表5中PTb-5-G平均吞吐量的测量函数和方法	>= 10 GB/s
	网络时延	等待对网络中的存储数据访问完成所引起的延时时间	使用GB/T 25000.23-2019表5中PTb-1-G平均响应时间的测量函数和方法	<0.5ms
	并发数	单位时间内，平均连接用户数	使用GB/T 25000.23-2019表7中PCa-2-G用户访问量的测量函数和方法	>= 5000用户
维护性	恢复时间目标（RTO）	从事件发生到完成恢复产品或服务、活动或者资源之间的时间段。	使用GB/T 20988-2007 信息系统灾难恢复规范	< 6h
	异常检测耗时	从异常发生到收到告警所消耗的时间	使用GB/T 25000.23-2019表5中PTb-1-G平均响应时间的测量函数和方法	10s

7.2 不同重要等级和不同访问频率数据的存储要求

按重要性等级数据存储可用性要求见表2，按访问频率等级数据判定规则与请求处理延迟要求见表3。

表2　重要性等级数据存储可用性要求

数据重要性等级	存储安全	可用性
核心数据	端到端加密，多重备份，保证100%不丢失	≥99.9999%
重要数据	访问审计，定期加密备份，保证100%不丢失	≥99.999%
一般数据	基础访问控制，周期性备份，保证100%不丢失	≥99.99%

表3　访问频率等级数据判定规则与请求处理延迟要求

数据访问频率等级	等级判定规则	请求处理延迟
热数据	数据访问频次从高到低，排序前10%部分数据	≤5ms
温数据	数据访问频次从高到低，排序前10%～30%部分数据	≤100ms
冷数据	数据访问频次从高到低，排序30%～100%部分数据	≤1min