# 子午工程空间天气科学数据统一存储与高效检索方法设计

李亚[1,2]　钟鼎坤[1]

1(中国科学院国家空间科学中心 太阳活动与空间天气全国重点实验室 北京 100190)
2(中国科学院大学 北京 100049）

**摘　要**　面向子午工程空间天气科学数据海量小文件管理，以及原始文件回溯访问与记录级条件检索并存的访问需求，提出一种统一存储与检索方法。该方法采用对象存储承载列式容器，按照数据产品类别与时间分区组织离散文件，在同一容器内协同保存原始文件内容与解析后的结构化记录，并结合元数据索引与容器内部统计裁剪机制，统一支撑文件回溯访问服务和记录条件检索服务。基于时间跨度 4 年、总规模约 2.4 TB、记录总数约 151 亿条的真实观测数据开展验证。结果表明，总体压缩率为 29.9%，其中文本类数据压缩率为 46.2%；可解析数值型数据由约 200 万个文件收敛为 93 个容器，单位空间记录承载能力提升约 80%。在并发访问条件下，文件回溯访问服务和记录条件检索服务均能稳定完成请求处理。该方法可有效降低海量离散小文件带来的文件数量膨胀程度，减少元数据维护、备份迁移和复制同步等方面的管理开销，为同类科学观测数据的统一组织与在线服务提供参考。

**关键词**　子午工程；空间天气科学数据；小文件治理；列式存储；统一存储与检索
**中图分类号**　（TP399）

最新录用

# Unified Storage and Efficient Retrieval Method for Space Weather Scientific Data of the Chinese Meridian Project

LI Ya[1,2]　ZHONG Dingkun[1]

1 (State Key Laboratory of Solar Activity and Space Weather, National Space Science Center, Chinese Academy of Sciences, Beijing 100190)
2 (University of Chinese Academy of Sciences, Beijing 100049)

**Abstract**　A unified storage and retrieval method is proposed for managing the massive number of small files in the Chinese Meridian Project space weather scientific data system and for supporting the coexistence of raw-file trace-back access and record-level conditional retrieval. In this method, columnar containers are hosted on object storage, and discrete files are organized according to data product category and time partition. Raw file contents and parsed structured records are collaboratively preserved within the same container, while file trace-back access and record-level conditional retrieval are jointly supported through metadata indexing and statistical pruning inside containers. Validation was conducted using real observational data spanning four years, with a total volume of approximately 2.4 TB and about 15.1 billion records. The results show that an overall compression ratio of 29.9% was achieved, including a compression ratio of 46.2% for text data. In addition, approximately 2 million files of parsable numerical data were converged into 93 containers, and the record-carrying capacity per unit storage space was improved by about 80%. Under concurrent access conditions, both file trace-back access services and record-level conditional retrieval services were stably completed. The proposed method can effectively alleviate file-number proliferation caused by massive discrete small files and reduce the management overhead associated with metadata maintenance, backup migration, and replica synchronization, thus providing a reference for the unified organization and online service of similar scientific observational data.

**Keywords**　Chinese meridian project; Space weather scientific data; Small-file management; Columnar storage;

Unified storage and retrieval

* 基金项目：国家重大科技基础设施子午工程项目资助
投稿日期：2026. 3. 14
E-mail: dkzhong@spaceweather.ac.cn

## 0 引言

为揭示日地空间环境多圈层间物质与能量耦合过程，我国先后建设了国家重大科技基础设施“空间环境地基综合监测链”（“子午工程”一期）与“空间环境地基综合监测网”（“子午工程”二期）[1-3]。随着子午工程二期观测能力提升、观测要素扩展及时间分辨率提高，系统持续产生多类型空间天气科学数据，数据规模快速增长，数据组织与服务压力不断增加。本文面向的研究对象为子午工程空间天气科学数据。该类数据在长期运行过程中表现出文件数量大、单文件体量小型化、数据格式异构并存以及多粒度访问需求并存等特征；类似问题在其他空间天气观测网络与地基综合观测系统中亦具有一定普遍性[4-8]。

子午工程是面向空间环境地基综合监测的国家重大科技基础设施，为科研群体和公众用户，提供监测数据与数据产品的推送、查询、下载、可视化及交互式处理等服务。与一期相比，二期观测范围进一步扩大，数据规模持续增长。当前数据服务仍以基于文件的获取方式为主，但随着数据体量不断增加，传统服务模式在海量数据场景下面临效率与支撑能力方面的挑战。基于此，有必要面向大规模数据应用需求探索更高效的检索、分析与共享支撑方式。

子午工程预计再运行十一年（一个太阳活动周期）以上，每年文件总数超过 3300 万个，年数据总量约为 7PB。其中，小于 10 MB 的文件占总文件数的 92.02%，而大于 1 GB 的文件仅占 2.66%，却贡献了 97.84% 的数据量。可见，数据文件整体呈现出海量小文件主导文件数量规模、少量大文件主导存储容量的分布特征。该特征使系统面临的主要压力不仅来自数据规模增长本身，更来自文件数量膨胀所带来的元数据管理、随机输入/输出（Input/Output，I/O）、备份迁移、复制同步及容灾切换等方面的运行负担。

在数据访问方面，传统空间物理研究通常以原始文件为基本单元，通过下载完整文件开展离线分析。而数据规模扩大后采用下载完整数据文件到本地再分析的传统方式会影响数据利用效率，系统开始需要支持基于时间范围、物理量选择与条件过滤的记录级检索能力以更方便进行远程数据分析[9-12]。

现有研究已在科学数据归档、元数据描述和结构化访问等方面形成一定基础。面向国际科学计划的数据组织与信息学研究表明，统一的元数据组织和服务框架是支撑分布式数据发现、共享与互操作的重要基础[13]。在具体实践中，IUGONET、ESPAS 等数据门户互联方案通过统一描述和语义关联，实现了跨平台数据资源的发现与访问[14]。面向日球物理和空间天气数据服务的近期研究进一步指出，统一的信息架构、元数据组织和服务机制是提升数据共享与访问效率的重要支撑；同时，围绕空间科学数据后续分析与智能应用的研究也进一步强调了数据规范化组织和高质量可访问性的重要性[15]。除数据发现、共享和文件级访问外，面向时间范围和参数选择的记录级访问也逐渐受到重视，如 HAPI 为日球物理时间序列数据提供了标准化访问方式[16]。

当前，子午工程数据中心针对原始文件访问与记录级访问并存的需求，采用原始文件与数据记录分离组织的方式：原始文件保存在文件库中，通过文件索引定位后完成文件访问；数据记录由原始文件解析、抽取后进入独立记录库，以支撑记录级访问。该方式在现有数据服务体系下能够分别满足原始文件获取与记录级访问两类需求。

但在长期运行条件下，这种分离式组织方式的局限也逐渐显现。一方面，海量离散文件持续累积，文件数量不断增长，增加了元数据维护、备份迁移和复制同步等方面的管理开销；另一方面，原始文件与解析记录分属不同存储体系，文件库与记录库需要长期协同维护，一致性控制与更新同步过程较为复杂。并且，文件级访问与记录级访问仍需依赖分离路径实现。子午工程数据中心当前的数据记录访问链路还需要依赖 Elasticsearch、HBase 及其相关基础环境共同支撑，其中 HBase 依赖 Hadoop 体系，导致系统部署依赖多、组件链条长，运行维护和故障排查成本较高，整体系统复杂度进一步上升。

在此基础上，本文进一步研究一种统一存储与检索方法，在缓解海量小文件问题的同时，保留原始文件内容，并在同一承载体系中协同组织原始文件与解析后的数据记录。相较现有分离式组织方式，该方法减少了系统依赖组件，降低了

整体系统复杂度及相应运维成本，并能够统一支撑文件回溯访问和记录条件检索两类服务。

## 1　子午工程空间天气数据存储与检索方法设计

### 1.1　总体设计

针对子午工程空间天气科学数据在长期运行条件下形成的海量小文件管理压力，以及文件访问与记录检索并存的服务需求，本文构建统一存储与检索框架。

子午工程数据体系涉及观测实体、原始数据文件、数据记录、数据产品和数据集等多个概念。观测实体表征台站、观测点和监测设备等数据来源；原始数据文件是设备按设备预定格式生成并保存的第一手监测数据实体，是数据入库与文件访问的基本载体；数据记录是从原始数据文件中抽取形成的数据单元，通常包含时间戳、观测参数及相关物理量；数据产品是监测数据，泛指监测设备产生的数据及经过后续各种处理形成的所有数据。数据集是按一定规则组织的数据文件或数据记录。

在此基础上，本文构建的统一存储与检索框架由元数据库、容器与数据承载三部分组成。整体设计如图 1 所示。

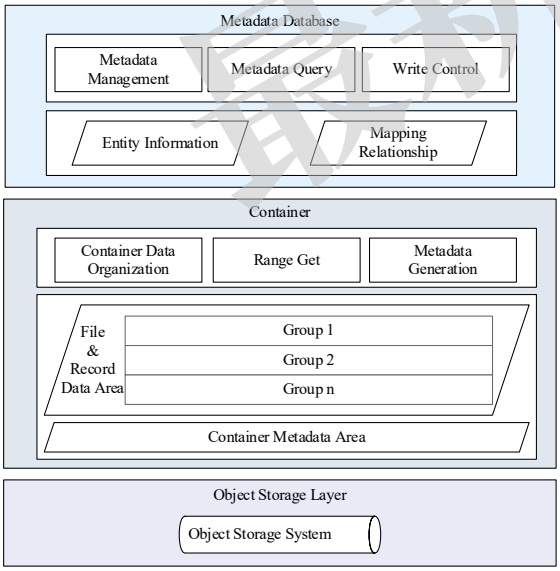



图 1　空间天气科学数据统一存储总体设计图

Fig.1　Overall Architecture of Unified Storage for Space Weather Scientific Data

其中，元数据库（Metadata Database）负责维护实体信息（Entity Information）及映射关系（Mapping Relationship），为写入控制与查询范围确定提供控制信息；容器（Container）在容器数据区（Container Data Area）中按组（Group）组织文件与记录数据（File & Record Data Area），实现原始文件与结构化记录的协同存储，并结合范围读取（Range Get）和元数据生成（Metadata Generation）支撑按需访问；对象存储层（Object Storage Layer）通过对象存储系统（Object Storage System）保存容器，提供持久化存储、扩展与范围读取能力。

该框架将数据组织规则与物理数据承载解耦。数据组织规则由元数据库统一描述，数据文件由对象存储系统保存，系统不再依赖路径命名、目录层级或应用约定表达主要数据语义。与此同时，原始文件内容与结构化记录不再分别落入相互独立的存储体系，而是在同一容器内统一组织，使文件回溯访问与记录级检索共享同一数据承载基础。由此，系统形成了元数据库负责范围收敛、容器内部结构负责精确读取的分层机制，为后续写入组织与检索裁剪提供基础。

### 1.2　元数据模型设计

元数据模型用于描述观测实体、数据产品结构与容器组织边界之间的关系，是系统实现写入控制与候选范围确定的关键控制信息。模型设计遵循语义信息外置、物理承载独立的原则，将数据来源、产品定义、字段结构与容器组织规则显式建模，使系统能够在不依赖存储路径和文件命名规则的条件下完成数据定位与访问。

模型由观测实体域、数据结构域和承载组织域三类语义域与其之间的关系构成，具体如图 2 所示。观测实体域描述站点、监测设备及其隶属关系，用于表达数据来源与归属边界；数据结构域描述数据产品类型、文件类型、字段及其类型约束，用于定义解析策略与记录结构；承载组织域描述数据产品在时间维上的分区规则、容器标识、存储位置及状态信息，用于确定容器组织边界并管理容器生命周期。

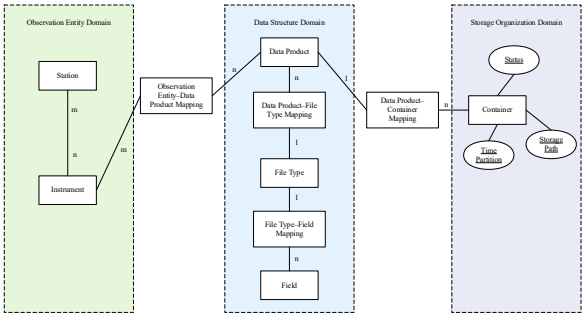



图 2　空间天气科学数据统一存储元数据模型图

Fig.2　Metadata Model of Unified Storage for Space

录用稿件

在关系组织上，模型维护四类核心映射关系：观测实体—数据产品映射、数据产品—文件类型映射、文件类型—字段映射，以及数据产品—容器映射。其中，观测实体—数据产品映射用于描述数据来源与产品归属，数据产品—文件类型映射与文件类型—字段集合映射共同定义原始文件到结构化记录的解析语义，数据产品—容器映射用于确定写入目标容器并支撑查询阶段的候选范围收敛。通过上述映射关系，系统建立了数据来源、数据产品结构与物理组织边界之间的统一关联。

在数据入库时，系统依据文件所属数据产品及其时间属性确定目标时间分区，并结合容器状态信息定位可写容器。当对应分区下不存在可写容器，或现有容器达到预设边界条件时，创建新容器并登记新的映射关系。该机制保证同一数据产品在同一时间窗口内的数据持续写入到组织规则一致的容器中，从而维持稳定的物理边界，并为查询阶段在容器粒度上的范围收敛提供基础。

在查询时，元数据模型承担候选容器范围确定功能。对于记录级检索，系统根据查询涉及的字段、产品类别和时间范围，定位相关数据产品及其覆盖分区，并返回候选容器集合；对于文件回溯访问，系统依据文件标识可关联的产品类别与观测时间信息，将请求收敛到少量候选容器。由于模型显式维护字段集合与数据产品之间的关系，系统不仅能够由字段条件反向推导相关产品类别，还可进一步确定相应的观测实体范围，从而简化查询参数构造的复杂度。

为控制元数据规模并降低维护成本，模型不维护文件标识到容器内部行组位置的细粒度映射。该类映射规模与文件数同阶增长，在长期运行条件下会带来显著的存储与更新开销。进一步的定位交由容器内部元数据完成，使外部元数据始终保持在容器粒度，在可维护性、可扩展性与定位效率之间取得平衡。

### 1.3 基于列式存储的容器设计

在容器划分策略上，系统采用“数据产品类别—时间分区”双维度组织规则。数据产品类别维度保证同一容器内字段结构、记录语义及解析规则保持一致，时间分区维度用于使同一数据产品在时间轴上保持连续并控制单个容器规模，从而增强时间范围查询场景下的范围收敛能力。两者共同构成容器边界控制的基础。

在此基础上，容器内部组织需同时支持按字段读取、压缩编码、范围获取与条件裁剪等要求。为此，本文采用列式存储思想组织容器内部数据，并引入面向范围读取的元数据描述机制，以兼顾原始文件回溯访问与记录级条件检索两类需求。

围绕这一目标，本文将“一个原始文件及其数据记录”定义为容器内部的一个组，并使该组与容器内部的行组级边界保持一致。采用这种组织方式，一方面可保持原始文件与其对应记录之间的天然关联，便于后续按组恢复文件内容；另一方面，也有利于以组为边界维护容器内部元数据，统一记录相关数据的结构信息、统计信息和位置描述信息，从而为范围读取、条件裁剪和目标定位提供支撑。基于上述设计，容器内部组织结构如图 3 所示。

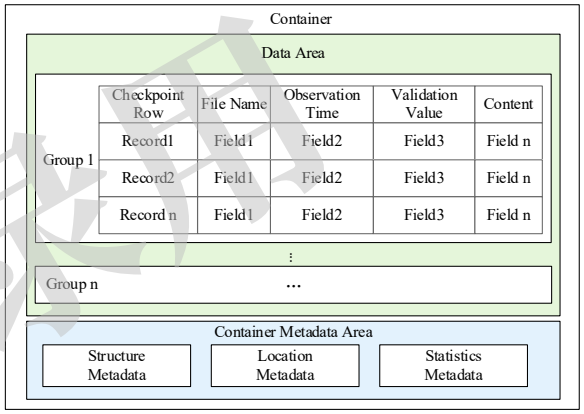



图 3　容器结构图

Fig.3　Container Structure

在单个组内部，组内第一条记录定义为锚点行，用于保存文件标识、观测时间、哈希校验值以及原始文件内容载荷等原始文件信息；其后各条记录定义为记录行，用于保存由该原始文件解析得到的结构化数据，具体如表 1 所示。锚点行与记录行共享统一表结构，但采用不同字段赋值方式：锚点行中的记录字段列以缺测值填充，不参与记录语义表达；记录行中的文件载荷字段以缺测值填充，不参与原始文件内容承载。通过这种方式，原始文件与解析记录在同一容器内实现统一组织，同时保留面向两类访问目标的差异化读取路径。

**表 1　容器内部数据统一表结构**

Tab.1　Unified Table Schema Inside the Container

| 字段名称 | 字段含义 | 适用行类型 | 锚点行 | 记录行 |
|---|---|---|---|---|
| file_name | 原始文件名 | 锚点行 | 赋值 | 空值 |

| | | | | |
|---|---|---|---|---|
| obs_time | 观测时间 | 锚点行 | 赋值 | 空值 |
| hash | 原始文件校验值 | 锚点行 | 赋值 | 空值 |
| content | 原始文件内容载荷 | 锚点行 | 赋值 | 空值 |
| field_1 | 解析字段 1 | 记录行 | 空值 | 赋值 |
| field_2 | 解析字段 2 | 记录行 | 空值 | 赋值 |
| field_n | 解析字段 n | 记录行 | 空值 | 赋值 |

对于文件回溯访问，可先定位目标组，再通过该组中的锚点行恢复原始文件内容，并结合校验信息验证一致性；对于记录级条件检索，则仅需访问参与条件判断和结果返回的必要字段，而无须读取原始文件内容载荷，从而避免文件内容对记录检索带来额外解码与传输负担。由此，锚点行承担文件可回溯保存功能，记录行承担结构化记录表达功能，而统一表结构则为二者在同一容器内的协同组织提供了可管理、可裁剪的物理基础。

为进一步缩小容器内部的读取范围，本文在容器尾部维护三类元数据信息。第一类为结构描述信息，用于描述容器统一表结构中的字段组成、字段类型及语义解释关系；第二类为组级位置描述信息，用于记录容器内各组的边界以及相关字段数据在容器中的位置与大小，从而支持目标组定位与按需读取；第三类为字段级统计信息，主要包括最小值、最大值、空值计数等，用于在记录级检索时预先判断某一组是否可能命中查询条件。基于这些信息，系统可先跳过明显无关的组，再仅读取命中组中参与条件判断和结果返回的必要字段，由此将读取范围由整个容器逐步收敛为相关组中的相关字段，避免全容器扫描与整体下载。

通过上述设计，容器内部形成了面向原始文件与解析记录协同组织的统一列式结构。组级边界保证了原始文件的独立映射关系与容器内范围收敛能力，字段独立组织保证了记录级检索时的按需读取能力，尾部元数据则为组定位、条件裁剪与范围读取提供了必要支撑。由此，系统在保持原始文件完整可回溯的同时，也能够兼顾面向条件检索场景的读取效率与存储压缩效果。

### 1.4 检索流程

为兼顾检索效率与元数据可维护性，本文采用“外部元数据范围收敛—容器内元数据裁剪”的检索与获取流程，具体如图 4 所示。该机制不依赖面向全局文件或记录的外部细粒度位置索引，而是先在外部完成候选容器收敛，再借助容器内部元数据完成进一步定位与裁剪。

对于记录级检索，系统首先根据查询字段、产品类别及时间条件，由元数据库确定相关数据产品、时间分区及候选容器集合。此阶段的目标是尽可能缩小访问范围，使后续读取不再面向全局数据空间，而仅针对有限容器集合展开。随后，系统读取候选容器尾部，利用元数据判断各组是否可能满足时间范围或字段谓词条件，跳过明显无关组；对于命中组，仅读取条件列和结果列对应的列块，并完成记录重建与结果返回。

对于文件回溯访问，系统首先依据文件标识可关联的产品类别与观测时间信息，将请求定位到候选容器；随后读取候选容器尾部，结合元数据对容器内部组进行进一步定位，并在目标组的锚点行中读取文件载荷与校验信息，恢复原始文件内容并返回结果。

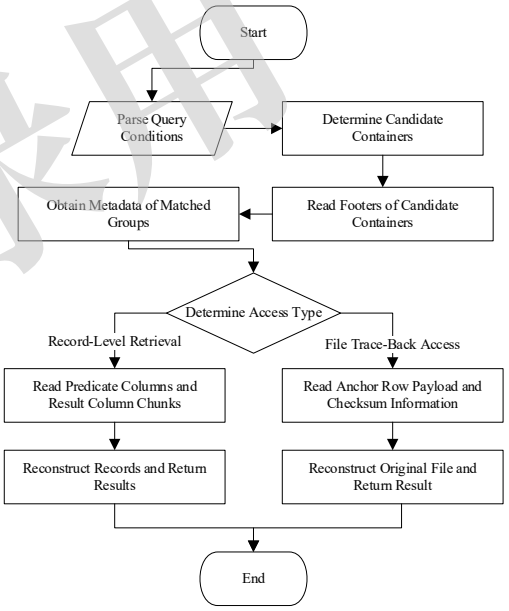



图 4　文件回溯访问与记录条件检索流程图

Fig.4　Workflow of File Trace-Back Access and Conditional Record Query

两类访问共享统一的候选容器收敛过程，差异主要体现在容器内部的目标读取对象不同：前者面向记录字段列块，后者面向锚点行载荷。通过上述协同机制，系统在保持外部元数据容器粒度可控的同时，实现了容器内部的组级与列级裁剪，能够统一支撑文件回溯访问与记录条件检索，从而形成面向海量小文件场景的统一存储与检索机制。

## 2　实验与分析

本章基于子午工程真实观测数据开展实验，

最终录用

从存储规模变化、文件数量控制、单位空间承载能力与并发访问性能四个方面验证所提方法的有效性。实验部分除考察本文方案自身的组织与服务能力外，还引入子午工程数据中心的数据系统现有组织方式作为基线方案，分析两种方案在文件级访问和记录级访问场景下的差异，以更直观地说明本文方法的改进效果。

### 2.1 实验环境与数据

实验采用 MinIO 对象存储集群作为统一数据承载，用于保存容器，并提供面向大规模数据的持久化存储、弹性扩展与范围读取能力。容器文件格式采用 Parquet，以满足统一表结构组织、组级边界保持、字段按需读取及尾部元数据辅助裁剪等设计需求。本文提出的方法面向统一存储与检索机制，本身不依赖于特定产品；在满足对象访问接口与容器内元数据裁剪能力要求的条件下，也可由其他对象存储系统与文件格式实现。实验访问请求由本地机器发起，由 Java 服务完成容器定位、读取调度与结果返回，并通过并发访问对系统性能进行测试与统计。

本文将子午工程数据中心的数据系统现有组织方式作为基线方案。基线方案采用文件库与数据记录库分离的组织方式：原始文件保存在对象存储中，通过文件索引定位目标文件位置后完成文件级访问；数据记录由原始文件抽取后进入独立的数据记录库，用于支撑记录级检索。相比之下，本文方案在统一承载体系中协同组织原始文件与解析后的结构化记录，并通过统一访问链路同时支撑文件回溯访问与记录条件检索。

为保证对比的可比性，两种方案在测试对象、服务场景、请求方式、接口功能、软件环境配置及统计指标等方面尽量保持一致，均围绕文件级访问与记录级检索两类服务开展并发性能测试。

后续并发性能实验中，基线方案运行于子午工程数据中心的生产环境，为尽量避免受业务影响，测试时间选在用户极少访问数据网站的凌晨 1 点左右；为了避免影响数据中心现有业务，本文方案在独立实验平台上完成测试。两个环境硬件条件并不一致，测试基线方案的数据中心生产环境的硬件规格高于本文方案的实验平台，相关结果主要用于说明本文方案的可用性及相对改进效果。

子午工程数据中心的生产环境包含一套 64 节点的 Kubernetes 容器云集群，一套 67 台存储服务器构成的 HBase 集群以及一套磁带库，所有节点通过万兆以太网互联，服务器 CPU 为 Intel Xeon Gold 5215 CPU @ 2.50GHz，每节点 128GB 内存，为 2019 年的主流硬件。

用于本文方案测试的独立实验平台的对象存储环境为 9 节点 MinIO 集群，总容量 288 TB，处理器为 Intel Xeon E5-2650 v4 @ 2.20 GHz，节点间通过千兆以太网互联，每节点 64GB 内存，为 2016 年的主流硬件。

本地服务与测试环境为 Windows 11 专业版 25H2，处理器 Intel Core i7-10700 @ 2.90 GHz，内存 32 GB；运行环境包括 JDK 1.8 与 Python 3.12，服务端采用 Spring Boot 2.3.7 实现数据访问接口。

实验数据来自子午工程真实观测数据，时间跨度为 4 年，共包含 46 种数据产品类型，文件数量约 370 万个，数据总规模约 2.4 TB，其中可解析数值型数据记录总数约为 151 亿条。从文件大小分布看，数据整体以 10 MB 以下中小文件为主，其中包含大量 5 MB 以下及 1 MB 以下文件；同时也包含一定数量的大文件。总体上，所选样本在产品类型、时间跨度、数据规模和文件大小分布等方面具有较好的代表性，能够较为真实地反映子午工程空间天气科学数据的实际存储与访问特征。

统计与分析指标包括：存储规模变化与压缩率、文件/容器数量变化、单位空间记录密度，以及并发访问的吞吐与延迟。压缩率定义为“节省空间大小 / 原空间大小”；记录密度定义为“记录数 / 容器大小”。

### 2.2 样本数据存储效果分析

基于 4 年观测样本数据的实验结果如表 2 所示。总体上，封装后总体压缩率为 29.9%。分类型来看，空间收益主要由文本类数据贡献，压缩率为 46.2%；而 PNG 与 JPG 图像数据的压缩率分别为 0.18% 和 2.5%，存储规模变化较小。结果表明，在 TB 级样本数据规模条件下，本文方案能够获得较稳定的空间收敛效果。

其中文本类数据的空间收敛效果最为明显，而图像类数据由于本身已采用专用压缩编码格式，进一步压缩空间有限。需要说明的是，数值型文本数据在空间压缩方面的收益，主要来自列式存储格式及其对应的编码压缩机制，这属于列式存储的通用能力，Parquet 在本文中只是选用的一种列式容器实现形式。本文方案的主要作用在于将大量离散文件收敛为少量容器，并在此基

础上实现原始文件与结构化记录的统一组织。

**表 2　样本数据总体实验结果表**

Tab.2　Experimental Results on the Sample Dataset

| 数据类型 | 原始存储规模/GB | 封装后存储规模/GB | 减少量/GB | 压缩率 |
|---|---|---|---|---|
| 文本文件 | 1 579 | 850 | 729 | 46.2 % |
| PNG 图像 | 560 | 559 | 1 | 0.18 % |
| JPG 图像 | 325 | 317 | 8 | 2.5 % |
| 总计 | 2 464 | 1 726 | 738 | 29.9 % |

进一步看，可解析数值型观测数据的存储效果如表 3 所示。与分离方案相比，统一组织方案在保持数据记录总数不变的前提下，实现了明显的空间收敛，并使单位空间记录承载能力提升约80%。

在文件规模方面，原先的海量离散文件被聚合为少量容器，实现了文件数量的数量级控制；在单位空间承载能力方面，存储密度提高约80%。该结果说明，方法能够在保持记录规模与可追溯性的同时，降低海量中小文件带来的文件与元数据管理压力，并通过更高的记录密度降低同等数据规模下的存储与传输负担。

**表 3　可解析数值型观测数据实验结果表**

Tab.3　Experimental Results for Numerical Observation Data

| 指标 | 分离方案 | 统一组织方案 |
|---|---|---|
| 数据规模/GB | 1579 | 850 |
| 文件/容器数量 | 200 万个文件 | 93 个容器 |
| 数据记录总数 | 151 亿条 | 151 亿条 |
| 存储密度/（条/MB） | 9 353 | 17 377 |

表 3 中“分离方案”指子午工程数据中心数据系统当前将原始文件以离散文件形式保存、并按既有方式管理的组织形式，未对文本内容进行统一压缩封装；“统一组织方案”是本文提出的统一存储与检索方案，即在同一列式容器内协同组织原始文件与解析记录的存储方式。

图像类数据以文件保存与获取为主，其实验结果如表 4 所示。总体上，图像类数据封装前后存储规模变化较小，其额外压缩收益有限；但文件数量实现了显著收敛，表明该类数据的主要收益体现在大量离散小文件聚合为少量容器，实现文件数量的数量级控制。可降低备份迁移、复制同步和归档处理等运维操作中的小文件管理开销，从而减少海量小文件场景下文件寻址、元数据维护和批量操作过程中的系统开销。

**表 4　图像类数据实验结果**

Tab.4　Experimental Results for Image Data

| 指标 | 分离方案 | 统一组织方案 |
|---|---|---|
| 总数据规模/GB | 885 | 876 |
| 总文件/容器数量 | 169 万个文件 | 42 个容器 |
| PNG 数据规模/GB | 560 | 559 |
| PNG 文件/容器数量 | 65 万个文件 | 29 个容器 |
| JPG 数据规模/GB | 325 | 317 |
| JPG 文件/容器数量 | 104 万个文件 | 13 个容器 |

综合来看，该组实验验证了本文方案在不同类型数据上的存储效果。对于可解析数值型数据，优势主要体现在空间收敛与组织效率提升；对于图像类数据，优势主要体现在文件数量收敛。总体上，本文方案改善了海量离散文件场景下的数据组织效率。

### 2.3　文件回溯访问服务性能测试

为评估在线服务场景下的可用性，并分析本文方法相对于现有组织方式的改进效果，本节开展文件回溯访问服务的并发性能实验。实验选取子午工程数据中心数据系统的现有组织方式作为基线方案，与本文方法进行对比。

测试设置 1、5、10、20、30、50、80、100 共 8 个并发梯度，并统计系统吞吐量（Queries Per Second，QPS）、平均延迟（Average Latency，Avg Latency）、P99 延迟（99th Percentile Latency，P99 Latency）等指标。由于本系统主要面向空间天气方向的科研用户，用户规模较小，实际业务场景下的并发访问需求相对有限，因此实验并发梯度整体设置在较低范围内。为分析文件体量差异对性能的影响，实验选取两类文本文件作为测试对象，分别为中型文本文件（42 MB）和小型文本文件（4.2 MB），结果分别如表 5 与表 6 所示。

从对比结果看，本文方案在文件回溯访问场景下的性能收益具有明显的文件规模差异。总体上，小型文件场景下改进更为明显，中型文件场景下改进相对有限。

对于小型文件场景，两种方案在低并发条件下差异较小，但随着并发数增加，本文方案的性能优势逐步显现。在并发数为 30 及以上的测试区间内，本文方案的吞吐量达到分离方案的 1.25～

最新录用

1.94 倍，同时 P99 延迟显著下降，说明其在小文件并发访问场景下具有更好的吞吐能力和尾部延迟控制效果。

对于中型文件场景，本文方案在低并发条件下表现出一定优势，但随着并发数继续增加，两种方案之间的差距逐渐减小。在较高并发区间，本文方案的吞吐量和 P99 延迟虽仍略优于分离方案，但改进幅度明显小于小型文件场景。

本文方案在文件回溯访问时还需经过容器内定位、数据提取和解压恢复等过程，而分离方案中的原始文件以离散形式直接保存，无需额外解压即可返回，这也在一定程度上削弱了本文方案在文件场景下的性能优势。

综合来看，本文方案的关键优势不在于文件回溯访问性能在所有场景下均显著提升，而在于在大幅减少小文件数量、完成统一组织的同时，仍能保持与分离方案相当的访问性能。其中，小型文件场景下还表现出一定的并发访问优势。这表明，本文方法在实现小文件收敛与统一组织目标的同时，未对文件回溯访问能力造成明显影响，能够稳定支撑文件回溯访问服务。

**表 5　中型文件场景并发读取性能表**

Tab.5　Concurrent File Read Performance in the Medium-File Scenario

| 并发数 | QPS | | Avg Latency /ms | | P99 Latency /ms | |
|---|---|---|---|---|---|---|
| | 分离方案 | 统一组织方案 | 分离方案 | 统一组织方案 | 分离方案 | 统一组织方案 |
| 1 | 1.49 | 1.31 | 673.00 | 761.00 | 673.00 | 761.00 |
| 5 | 2.15 | 2.81 | 1518.40 | 848.40 | 1612.00 | 869.00 |
| 10 | 2.38 | 3.18 | 2752.70 | 2235.20 | 3536.00 | 2609.00 |
| 20 | 2.35 | 2.59 | 5403.10 | 6306.25 | 7716.00 | 6870.00 |
| 30 | 2.40 | 2.76 | 7426.07 | 8159.97 | 11631.00 | 10027.00 |
| 50 | 2.45 | 2.46 | 11425.62 | 13926.20 | 19544.00 | 19421.00 |
| 80 | 2.46 | 2.49 | 17619.28 | 20175.02 | 31714.00 | 31824.00 |
| 100 | 2.48 | 2.68 | 21668.88 | 22928.59 | 40016.24 | 36404.74 |

**表 6　小型文件场景并发读取性能表**

Tab.6　Concurrent File Read Performance in the Small-File Scenario

| 并发数 | QPS | | Avg Latency /ms | | P99 Latency /ms | |
|---|---|---|---|---|---|---|
| | 分离方案 | 统一组织方案 | 分离方案 | 统一组织方案 | 分离方案 | 统一组织方案 |
| 1 | 8.13 | 9.26 | 123.00 | 108.00 | 123.00 | 108.00 |
| 5 | 5.44 | 4.94 | 79.20 | 86.60 | 127.00 | 115.00 |
| 10 | 10.94 | 9.88 | 107.10 | 91.20 | 172.00 | 129.00 |
| 20 | 21.72 | 19.34 | 169.35 | 100.15 | 334.00 | 137.00 |
| 30 | 22.64 | 28.33 | 454.43 | 120.37 | 935.00 | 161.00 |
| 50 | 22.75 | 38.05 | 989.02 | 339.78 | 1517.00 | 511.00 |
| 80 | 23.15 | 41.64 | 1690.60 | 723.27 | 2691.00 | 1076.00 |
| 100 | 23.19 | 44.88 | 2153.76 | 882.81 | 3650.19 | 1644.85 |

### 2.4　条件检索记录服务性能测试

在文件回溯访问服务测试的基础上，本节进一步评估记录条件检索服务的并发处理能力。记录条件检索测试选取一类按天分割的数据产品作

最新录用

为对象。测试过程中，每次请求均以 1 天时间范围作为检索窗口，并从该数据产品字段中随机选取两个物理量构成组合条件，对满足条件的记录进行筛选与返回。测试设置 1、5、10、20、30、50、80、100 共 8 个并发梯度，将子午工程现有记录访问方式作为基线方案，与本文方法进行对比，实验结果如表 7 所示。

**表 7　数据记录并发检索性能表**

Tab.7　Concurrent Record Query Performance

| 并发数 | QPS | | Avg Latency /ms | | P99 Latency /ms | |
|---|---|---|---|---|---|---|
| | 分离方案 | 统一组织方案 | 分离方案 | 统一组织方案 | 分离方案 | 统一组织方案 |
| 1 | 0.02 | 5.75 | 44065.00 | 174.00 | 44065.00 | 174.00 |
| 5 | 0.04 | 4.63 | 107129.60 | 156.00 | 115977.00 | 185.00 |
| 10 | 0.06 | 9.24 | 146515.50 | 155.80 | 157611.00 | 181.00 |
| 20 | 0.07 | 18.20 | 291466.75 | 169.10 | 305918.00 | 196.00 |
| 30 | 0.06 | 27.20 | 471793.77 | 216.67 | 477089.00 | 258.00 |
| 50 | 0.06 | 36.28 | 820069.94 | 497.96 | 838001.00 | 740.00 |
| 80 | 超时未完成 | 26.95 | 超时未完成 | 1640.17 | 超时未完成 | 2685.00 |
| 100 | 超时未完成 | 25.93 | 超时未完成 | 1929.16 | 超时未完成 | 3396.97 |

从对比结果看，记录条件检索是本文方案改进最显著的场景。总体上，本文方案在各并发梯度下均表现出明显优于基线方案的吞吐能力和延迟表现，且在高并发条件下仍能保持请求成功。

从变化趋势看，基线方案在并发增加后吞吐量提升极为有限，而平均延迟和尾部延迟持续快速升高；相比之下，本文方案的吞吐量能够随并发增加而明显提升，延迟增长幅度相对可控。在可完成对比的并发区间内，本文方案的吞吐量相较基线方案已呈现数量级差异，P99 延迟也由基线方案的高秒级显著降低至毫秒级，说明两种方案在记录级检索场景下存在明显性能差异。

造成这一差异的原因，主要与记录检索的数据组织方式和访问路径有关。本文方案基于统一承载结构组织记录数据，结合列式存储、对象存储和范围读取机制，仅访问与查询条件相关的字段和数据区段，因此能够减少无关数据扫描开销。相比之下，基线方案采用原始文件与记录库分离的组织方式，记录级访问需要依赖独立记录库完成条件筛选；在查询范围扩大、并发增加时，记录库的筛选与返回开销更易累积，因此更容易出现吞吐受限和延迟升高。

综合来看，记录条件检索是本文方案优势最为集中的应用场景。实验结果表明，统一组织方式不仅能够支持记录级条件检索服务，而且相较现有分离式组织方式，在并发访问条件下表现出更高的检索效率和更好的服务稳定性。

### 2.5　工程指标对比验证

为进一步评估本文方法对工程应用需求的支撑能力，本节开展了数据文件索引查询响应测试。该测试面向网页数据服务场景，以单次查询 1000 条数据文件索引作为测试场景，统计请求的平均响应时间与最大响应时间。测试结果表明，现有系统在该规模下的平均响应时间为 47.43 ms，最大响应时间为 129 ms；本文方案的平均响应时间为 54.19 ms，最大响应时间为 98 ms。两种方案均显著优于 1 s 的设计要求。总体来看，两种方案在查询 1000 条数据文件索引时均能实现稳定的亚秒级响应，能够满足网页侧快速索引访问需求。

子午工程二期在初步设计时对数据服务能力提出了数据下载服务能力、平台并发访问规模以及网页查询响应速度等要求。为进一步说明本文方案与现有系统及工程指标之间的关系，表 8 给出了相关对照结果。由表可见，本文方案在各项指标上均满足子午工程二期提出的工程要求。其中文件下载服务能力方面，在 4.2 MB 小文件、100 并发场景下，本文方案的平均延迟和 P99 延迟均低于现有系统，说明在该类服务场景下本文方案具有更好的访问性能；平台并发访问能力方面，本文方案已在文件下载与数据记录访问场景下覆盖相应并发量级；数据查询响应速度方面，

最新录用

两种方案均满足小于 1 s 的设计要求。

**表 8　工程设计指标对比表**

Tab.8　Comparison with Engineering Design Indicators

| 场景 | 设计指标 | 现有系统结果 | 本文方案结果 |
|---|---|---|---|
| 数据下载服务能力（5 MB） | 支持 100 个并发用户，且响应时间小于 2 分钟 | 在 4.2 MB 小文件、100 并发场景下，Avg Latency 为 2153.76 ms，P99 Latency 为 3650.19 ms | 在 4.2 MB 小文件、100 并发场景下，Avg Latency 为 882.81 ms，P99 Latency 为 1644.85 ms |
| 数据服务平台并发访问数 | 支持 300 个用户同时访问数据服务平台，按 20% 用户同时在线计 | 相关文件下载服务已覆盖该并发量级 | 文件下载与数据记录访问服务均已覆盖该并发量级 |
| 数据查询响应速度 | 查询 1000 条数据文件索引，响应时间小于 1 s | 平均响应时间为 47.43 ms，最大响应时间为 129 ms，优于指标要求 | 平均响应时间为 54.19 ms，最大响应时间为 98 ms，优于指标要求 |

综合来看，本文方法在数据下载服务能力、平台并发访问规模和数据查询响应速度等方面均满足子午工程二期提出的工程要求。结合前文文件回溯访问与记录条件检索的并发对比实验可以看出，本文方法在文件下载和数据记录访问等关键服务场景下相对于现有系统具有更好的性能表现；在数据文件索引查询响应速度这一指标上，现有系统与本文方案均已显著优于设计要求，其中本文方案在最大响应时间上更低。与此同时，本文方法通过统一组织原始文件与结构化记录，减少了分离式组织方式下对多类组件的依赖，降低了系统复杂度，并相应降低了系统运维成本与运维难度。总体而言，本文方法能够在满足工程部署需求的前提下，为实际数据服务系统建设提供支撑。

## 3　结论

本文研究了一种面向子午工程空间天气科学数据的统一存储与检索方法，可在同一体系内同时支撑文件回溯访问和记录条件检索。该方法以对象存储为统一承载，通过列式容器实现离散文件聚合，并在同一容器内协同组织原始文件与结构化记录。

实验结果表明，本文方法在保持原始文件可回溯能力的前提下，实现了存储空间收敛和文件数量显著减少。对于可解析数值型数据，单位存储空间的数据记录承载能力得到提高；对于图像类数据，虽然未带来明显额外压缩收益，但实现了文件数量收敛，且未引入显著额外存储开销。

并发访问测试表明，本文方法能够稳定支撑文件回溯访问服务和记录条件检索服务。其中，文件回溯访问场景下的性能收益主要体现在小型文件的中高并发访问中；记录条件检索场景下，相较现有分离式组织方式，本文方法表现出更高的检索效率和更稳定的服务能力。工程指标对比结果进一步表明，本文方法能够满足子午工程二期的数据服务能力要求。

总体来看，本文方法在满足工程应用需求的同时，减少了分离式组织方式下对多类存储与检索组件的依赖，降低了整体系统复杂度，并减轻了相应的运维成本与管理难度，为空间天气科学数据服务系统的统一组织、统一访问与长期演进提供了一种可行的实现路径。

## 参考文献（References）


[1] WANG C, WANG J Y, XU J Y. Research advances of the Chinese Meridian Project in 2020-2021[J]. *Chinese Journal of Space Science*, 2022, **42**(4): 574-583

[2] QIN Jiamei, ZHONG Dingkun, FENG Xueshang, *et al*. A summary of data resources produced by the Meridian Project[J]. *China Scientific Data*, 2021, **6**(2): 7-29 (秦佳媚, 钟鼎坤, 冯学尚, 等. 子午工程数据资源概述[J]. 中国科学数据, 2021, **6**(2): 7-29)

[3] WANG C, XU J Y, CHEN Z Q, *et al*. China's ground-based space environment monitoring network—Chinese Meridian Project (CMP)[J]. *Space Weather*, 2024, **22**(7): e2024SW003972

[4] MASSON A, FUNG S F, CAMPOREALE E, *et al*. Heliophysics and space weather information architecture and innovative solutions: current status and ways forward[J]. *Advances in Space Research*, 2024

[5] MAMPAEY B, DELOUILLE V, VANSINTJAN R. The SOLARNET Virtual Observatory: an effective way to search solar datasets and cross-search with solar events[J]. *Solar Physics*, 2025, **300**(2): 16

[6] SELLERS S G, SHETYE J, CHRISTIAN D J, *et al*. The Sunspot Solar Observatory Data Archive: continuing operations at the Dunn Solar Telescope[J]. *Solar Physics*, 2025, **300**(8): 103

[7] PACINI A A, RACHMELER L, BAUGH K. NOAA's national

centers for environmental information space weather data inventory status[J]. *Space Weather*, 2025, **23**(7): e2025SW004529

[8] ABE S, UMEMURA N, KOYAMA Y, *et al*. Progress of the IUGONET system-metadata database for upper atmosphere ground-based observation data[J]. *Earth, Planets and Space*, 2014, **66**(1): 133

[9] ANGELOPOULOS V, CRUCE P, DROZDOV A, *et al*. The space physics environment data analysis system (SPEDAS)[J]. *Space Science Reviews*, 2019, **215**(1): 9

[10] ROBERTS D A, THIEMAN J, GÉNOT V, *et al*. The SPASE data model: a metadata standard for registering, finding, accessing, and using heliophysics data obtained from observations and modeling[J]. *Space Weather*, 2018, **16**(12): 1899-1911

[11] FUNG S F, MASSON A, BARGATZE L F, *et al*. SPASE metadata as a building block of a heliophysics science-enabling framework[J]. *Advances in Space Research*, 2023, **72**(12): 5707-5752

[12] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, *et al*. The FAIR guiding principles for scientific data management and stewardship[J]. *Scientific Data*, 2016, **3**(1): 160018

[13] FOX P, KOZYRA J. eScience and informatics for international science programs[J]. *Progress in Earth and Planetary Science*, 2015, **2**(1): 12

[14] RITSCHEL B, BORCHERT F, KNEITSCHEL G, *et al*. Experiments using Semantic Web technologies to connect IUGONET, ESPAS and GFZ ISDC data portals[J]. *Earth, Planets and Space*, 2016, **68**(1): 181

[15] PODUVAL B, MCPHERRON R L, WALKER R, *et al*. AI-ready data in space science and solar physics: problems, mitigation and action plan[J]. *Frontiers in Astronomy and Space Sciences*, 2023, **10**: 1203598

[16] WEIGEL R S, VANDEGRIFF J, FADEN J, *et al*. HAPI: an API standard for accessing heliophysics time series data[J]. *Journal of Geophysical Research: Space Physics*, 2021, **126**(12): e2021JA029534

## 作者简介



**李亚** 男, 2000 年 4 月出生于安徽省六安市, 现为中国科学院国家空间科学中心 2023 级硕士研究生, 主要研究方向为空间智能信息处理与应用.
E-mail: liya_426@163.com



**钟鼎坤 （通信作者）** 男, 1979 年 8 月出生于广东省韶关市, 现为中国科学院国家空间科学中心副研究员, 硕士生导师,主要研究方向为空间天气科学数据管理和可视分析.
E-mail: dkzhong@spaceweather.ac.cn

最新录用