引文格式：王意. 面向混合存储的时空大数据存储策略研究[J]. 地理空间信息, 2023, 21(9): 10-13. doi:10.3969/j.issn.1672-4623.2023.09.003

面向混合存储的时空大数据存储策略研究

王　意[1]
（1. 山东省国土测绘院，山东 济南 250013）

摘　要：混合存储能综合各类管理系统和设备的优点，成为时空大数据存储的理想选择，但不同管理系统下的时空大数据之间难以建立有效关联，且数据存储策略缺少对时空大数据的动态感知和弹性调度能力。分析了当前典型混合存储技术产品，根据时空大数据特性，提出了面向混合存储的时空大数据存储策略。通过建立数据存储模型，在数据层面实现一个完整的时空大数据存储生态链，设计数据迁移函数来量化数据价值，灵活动态地匹配存储设备。测试表明，混合存储策略能有效提升时空大数据的存储访问性能，更好地服务于时空大数据的管理应用。

关键词：时空大数据；混合存储；存储模型；数据迁移
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2023）09-0010-04

Research on Spatio-temporal Big Data Storage Strategy for Hybrid Storage

WANG Yi[1]
(1. Shandong Provincial Land Surveying and Mapping Institute, Jinan 250013, China)

Abstract: The hybrid storage has become an ideal solution for spatio-temporal big data storage, which can fully take advantage of both hybrid storage management systems and hybrid storage devices. However, it is difficult to establish effective correlation between spatio-temporal big data under different management systems, and the data storage strategy can neither dynamically perceive spatio-temporal big data nor adjust the position of spatio-temporal big data in time. According to the characteristics of spatio-temporal big data, we proposed a storage strategy for hybrid storage by the analysis of current typical hybrid storage technology products. We realized a complete spatio-temporal big data storage ecological chain at the data level by building organization and storage model. Spatio-temporal big data could match storage devices flexibly and dynamically by designing data migration function to quantify data value. The test result shows that the hybrid storage strategy can effectively improve the storage access performance of spatio-temporal big data, and better serve the management and application of spatio-temporal big data.

Key words: spatio-temporal big data, hybrid storage, storage model, data migration

时空大数据是基于统一时空基准，活动于时空中与位置直接或间接关联的大数据[1]，是支撑宏观经济决策的战略级资源[2]。时空大数据存储性能的优劣直接影响其计算效率和服务性能[3]。目前混合存储是时空大数据管理中一种较好的选择[4-5]。本文在分析当前主流存储管理系统和硬件的基础上，按照数据类型和结构对存储对象进行了划分；结合时空混合索引方法，建立了时空大数据存储模型；分析了时空大数据存储影响因素，并进行定量化表达；提出了基于混合存储技术的时空大数据存储策略，建立了高效可靠的时空大数据放置与迁移方法，以期更好地服务于时空大数据存储和业务化应用。

1　现状综述

20 世纪 90 年代关系型数据库成为空间数据存储管理系统的主流应用模式。由于时空大数据非结构化的数据读写场景，分布式文件系统和分布式非关系型数据库技术应运而生（表 1）。

传统的机械硬盘（HDD）面对大数据时代的挑战已显得力不从心，新型存储介质具有体积小、能耗

表 1　混合存储中常见的存储管理模式

类型	典型产品	优势	不足
关系型数据库	MySQL	更兼容传统GIS管理系统	难以支撑非结构化数据的复杂空间分析
分布式文件系统	Hadoop	支撑高并发和并行计算	空间查询、统计分析能力较弱
非关系型数据库	MongoDB	扩展性好，可高效集成空间数据	与传统GIS存在差异，操作有局限性

收稿日期：2022-04-08；修回日期：2022-06-29。
项目来源：山东省“十三五”基础测绘规划基金资助项目（201605097）。
作者简介：王意（1985—），硕士，研究方向为地理信息基础设施建设与维护，E-mail：wy09land@163.com。

低、带宽高、时延低、抗震性强、可靠性高等特点。其中，闪存的发展最迅速，已达到了实用化水平，固态盘（SSD）是闪存最主要的形式。

2　混合存储策略

2.1　问题描述

1）不同管理系统间的数据关联问题。时空大数据在访问特性和时效要求上差异较大[6-8]。混合存储需满足不同管理模式下数据的动态关联需求，将通用的混合存储技术与空间信息技术（空间大数据组织方法、空间分布式索引等）深度融合，形成集中统一的数据访问接口。

2）数据存储策略缺少对时空大数据的动态感知和弹性调度能力。时空大数据的存储策略应具备动态感知和弹性调度数据的能力，可根据应用的历史访问记录、存储设备特征等因素实时调整数据存储路径；通过一个有效的迁移算法为不同数据动态选择存放介质，使整个存储系统达到存储效率和经济性的最佳平衡，从而最大程度地提升整个存储系统的性能。

2.2　总体设计

在混合存储架构中的数据层构建面向时空大数据的关系型数据库集群、NoSQL数据库集群、分布式文件系统集群的存储框架，实现不同管理模式下的数据动态关联；以SSD、HDD作为二级存储，根据I/O特性和数据价值，将不同类别的数据分配到SSD或HDD上；设计数据迁移算法，实现数据在不同介质间的移动。

2.3　数据存储模型构建

构建结构化数据、半结构化数据和非结构化数据3种存储模型，并建立元数据及其索引，将具有相同编码的数据进行关联。

1）将一个结构化的数据库或子数据库切分为若干具有固定大小的组块，并对数据块建立两层索引。存储模型包括数据块集合、第一层索引和第二层索引3个部分，其中数据块中包含空间几何信息、空间数据信息和其他属性信息（图1）。通过统一的空间编码方法和哈希编码方法使每条数据具有唯一编码。

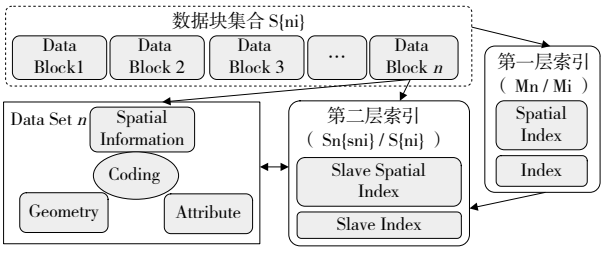


图1　结构化数据存储模型

2）半结构化数据通过面向数据片的方式管理和查询数据，建立空间编码（索引）。半结构化数据存储模型包括数据分片集合和空间编码两大部分，其中数据分片中包含数据、编码深度和编码（图2）。数据分片无需进行二次检索，通过统一的空间编码和解码即可快速查询和获取瓦片。

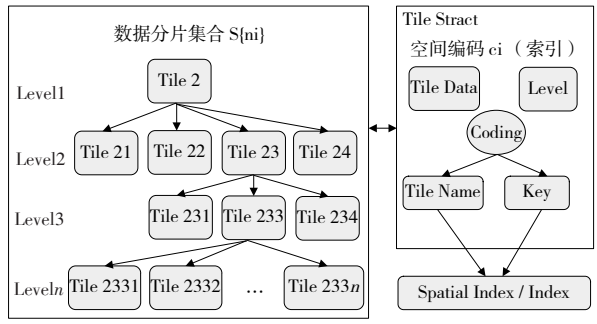


图2　半结构化数据存储模型

3）非结构化数据不能直接存储矢量数据，需转换为能保存的空间信息和属性信息。非结构化数据存储模型包括文件集合和关联编码两大部分，其中文件集合记录文件信息，并通过统一的关联编码方法使其具有唯一编码（图3）。关联编码属于第一级索引，对于常规索引检索的 n 个结果，可采用分布式并行的方法进行并行读取，并可通过aci对文件进行命名，与其他具有相同编码的数据进行关联。

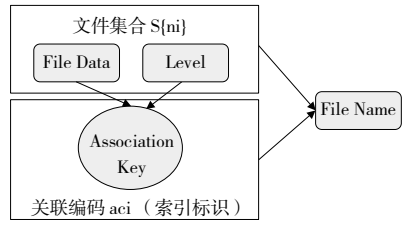


图3　非结构化数据存储模型

2.4　数据放置与迁移策略

影响因素主要包括：①空间分辨率，反映数据的空间详细程度，值越高，识别物体的能力越强；②时间分辨率，表示重复探测同一目标时，相邻两次探测的时间间隔；③光谱分辨率，反映成像的波段范围，划分越细、波段越多，值越高；④存储时间，数据存放较长时间后，其被访问频率将降低，数据价值也将降低；⑤数据热度是反映数据价值的直观指标，利用价值高的数据会被反复读取加以利用；⑥随机度，使SSD服务绝大多数的随机I/O访问，HDD服务剩余的I/O访问，匹配度越高的数据，越能体现或挖掘其数据价值；⑦数据大小，数据量越小，越有利于存储资源的利用，数据大小与价值成反比。

2.4.1　数据静态放置策略

本文选择时空大数据的空间分辨率、时间分辨

率、光谱分辨率作为评价指标，不随数据应用而改变。根据主观经验设置权重 Q，满足条件的数据优先使用性能更高的SSD作为存储介质。假设某应用可调用 n 个数据，单个数据价值的计算公式为：

$$V(S_i, T_i, F_i)=Q_s\times J_{s_i}+Q_t\times J_{t_i}+Q_f\times J_{f_i}$$

式中，Q_s、Q_t、Q_f 分别为空间分辨率、时间分辨率、光谱分辨率的权重，由人为确定；J_{s_i}、J_{t_i}、J_{f_i} 分别表示数据 i 的3个指标是否满足应用需求，若满足，值为1，不满足，值为0。

选择该应用价值最大的数据优先放置在SSD上，计算公式为：

$$V=\max(V(S_1, T_1, F_1), V(S_2, T_2, F_2), \cdots, V(S_n, T_n, F_n))$$

2.4.2 数据迁移函数

数据每次被访问的时间集为 $\{t_1, t_2, \cdots, t_n\}$，距离当前时刻 t 的时长为 $t-t_1, t-t_2, \cdots, t-t_n$，用 $C_1, C_2, \cdots, C_n$ 表示。设数据访问热度为 q，随机度为 r，初始值均为0，访问一次数据，q 加1，判断该访问的起始位置是否对应该数据上次访问的结束位置，若相同且与上次访问类型相同、时间间隔也在给定的阈值内，则 r 不变；否则 r 加1。在存储时间 $C_1, C_2, \cdots, C_n$ 内的访问热度为 $q_1, q_2, \cdots, q_n$，随机度为 $r_1, r_2, \ldots, r_n$，计算公式为：$C=\sum_{i=1}^{n}\frac{1}{C_i}, q=\sum_{j=1}^{n}q_j, r=\sum_{k=1}^{n}r_k$。由此可知，数据价值与 q、r 成正比，与 C、数据大小成反比。

单个数据价值和所有数据总价值的计算公式为：

$$V=\sum_{j=1}^{n}q_j\times\sum_{k=1}^{n}r_k\times\sum_{i=1}^{n}\frac{1}{C_i}\times\frac{1}{S}$$

$$Q=\sum_{i=1}^{n}V_i x_i$$

$$x_i=\begin{cases}1, \text{文件}i\text{存储在SSD上}\\0, \text{文件}i\text{存储在HDD上}\end{cases}$$

所有数据的总价值应尽可能的大，根据数据价值变化，系统I/O动态地在SSD、HDD之间进行数据迁移，以保证价值较高的数据优先使用SSD存储。设数据大小为 s_i，应保证集合中SSD上所有数据大小的总和不能超过SSD的总空间大小（$\text{CAPACITY}_{\text{SSD}}$），即

$$\sum_{i=1}^{n}x_i s_i\leqslant\text{CAPACITY}_{\text{SSD}}$$

设置模型的起止时间和SSD的空间总量，计算得到使数据总价值最大的存储方案，进而调整数据存储位置。

3 实验与性能分析

为验证本文提出的时空大数据存储策略，以山东省地理信息时空大数据中心为依托，分别测试数据存储和加载效率。实验设备配置见表2。

表2 实验设备配置

设备类型	数量/台	用途
混合存储管理端	1	部署数据存储框架、数据迁移
普通客户终端	1	调度、数据存储和读取
PostgreSQL集群	3	综合库管理系统、基础数据库管理系统等关系型数据库服务
MongoDB集群	6	统一的NoSQL存储服务，主体存储矢量/栅格瓦片
HDFS集群	6	文件、音频等数据分布式存储服务

3.1 实验数据

本文选取两组数据：①山东省省级基础测绘“十二五”地形要素数据，共5 977个MDB、51.7 GB，“十三五”数字正射影像6 578幅、2.73 TB；②1.6亿条矢量要素数据、120 TB影像数据和500 GB地形数据（DEM数据），数据具有明显的时空大数据特征。

3.2 存储时间

两组数据采用相同配置的终端，分别存入部署混合存储策略的数据库（实验库）和原始库，初始存放位置均为HDD。数据入库时间对比见表3，可以看出，第一组实验库的数据入库时间比原始库约减少10%，入库效率小幅提升，这是由于数据量较少时，存在较固定的数据传输和备份时间消耗，因此优势不明显；第二组实验库的数据入库时间缩短近一半，而矢量要素数据和地形数据入库时间缩短为原时长的1/8，说明本文设计的混合存储策略综合考虑了数据特征、混合存储管理系统、混合存储设备3个要素，提高了数据放置的合理性，减少了系统开销，大幅提高了入库效率。

表3 数据入库时间对比

实验组别	数据名称	原始库入库时间	实验库入库时间/h
第一组	地形要素数据	3 h20 min	3
	影像数据	4 h25 min	4
第二组	矢量要素数据	160 h	20
	影像数据	45 h	20
	地形数据	32 h	4

3.3 加载时间

本文对实验数据进行加载浏览，分别测试了不同数据存储策略下10万、100万、500万条数据的初始加载时间和多次打开数据库后的加载时间，结果见表4，可以看出，混合存储管理端初次加载时长有所减少，

表4　数据加载时间对比

数据量	普通终端加载时间		混合存储管理端加载时间	
	初次加载/s	第五次打开数据库/s	初次加载/s	第五次打开数据库/s
10万	3~5	3~5	3~5	≤3
100万	10~12	10~12	8~10	≤10
500万	30~35	30~35	20~30	≤15

并随数量的增加，加载时长减少越多（30%），说明混合存储策略数据加载时效率显著提升；普通终端多次打开数据库后的加载时间没有明显变化，而混合存储管理端在多次打开数据库后，加载时间约为初次加载时间的50%，说明通过数据迁移，将随机访问和热点访问的数据迁移到SSD上，有效减少了整个系统的访问延迟。

4　结　语

本文分析了混合存储技术在时空大数据存储中的应用现状，提出了面向混合存储的时空大数据存储组织模型和数据存储策略；并基于山东省地理信息时空大数据中心开展性能实验，对比了普通终端和混合存储策略终端的时空大数据存储和加载效率。结果表明，在存储数据量较少的传统测绘数据成果时，混合存储策略效率提升不明显；但随着数据量和数据种类的增加，混合存储策略的效率显著提升；在多次数据读取后，数据的动态迁移可有效提升数据访问性能。

参考文献

[1] 王家耀,武芳,郭建忠,等. 时空大数据面临的挑战与机遇[J]. 测绘科学,2017,42(7):1-7

[2] 刘先林. “互联网＋”时代GIS的智能特征及展望[J]. 测绘科学,2017,42(2):1-4

[3] 李德仁. 论“互联网+”天基信息服务[J]. 遥感学报,2016,20(5):708-715

[4] 李鸿飞,杜溢墨,曾熠,等. 异构混合存储的软硬件协同数据放置策略[J]. 国防科技大学学报,2021,42(2):64-71

[5] 罗保山,张鑫奘,王栩,等. 混合存储数据迁移策略研究[J]. 计算机技术与发展,2016,26(6):82-86

[6] 李仕峰. 多源异构数据时空融合关键技术研究与应用[J]. 地理空间信息,2021,19(10):19-21

[7] 刘善磊,潘九宝,李梦梦,等. 多源地理信息数据匹配质量评价研究[J]. 地理空间信息,2022,20(1):118-121

[8] 杜江毅,边馥苓. 面向大数据的空间数据挖掘综述[J]. 地理空间信息,2017,15(1):8-12

（上接第9页）

演变格局监测，后期将从数据补全补实的角度，加强黄河流域铁路交通联系时空演变格局分析、变化机理分析、格局优化模拟等方面研究，为黄河流域高质量发展提供更系统深入的决策支持。

参考文献

[1] 许玉姣. 新经济态势下黄河流域高质量发展现状与对策[J]. 甘肃科技纵横,2021,50(6):12-14

[2] 新华社. 中共中央国务院印发黄河流域生态保护和高质量发展规划纲要[J]. 中国水利,2021(21):3-16

[3] 胡昊宇,黄苇绒,李沛霖,等. 流空间视角下中国城市群网络结构特征比较:基于铁路客运班次的分析[J]. 地球信息科学学报,2022,24(8):1 525-1 540

[4] 齐昕,王立军,张家星,等. 高铁影响下城市群空间关联形态与经济增长效应研究[J]. 地理科学,2021,41(3):416-427

[5] Ge F, Chen W, Zeng Y, et al. The Nexus Between Urbanization and Traffic Accessibility in the Middle Reaches of the Yangtze River Urban Agglomerations, China[J]. International Journal of Environmental Research and Public Health,2021,18(7):3 828

[6] 王海江,苗长虹. 基于O-D网络的全国中心城市铁路客运联系及其分布规律[J]. 经济地理,2014,34(5):83-90

[7] 杨春华,吴晋峰,周芳如,等. 铁路通达性变化对区域旅游业的影响:以京津冀、长三角地区对比为例[J]. 经济地理,2018,38(2):188-196

[8] Ortega E, Monzón A, López E. The Influence of Spatial Data Allocation Procedures on Accessibility Results: The Case of High-speed Rail Networks[J]. Applied Geography,2018,94:241-250

[9] Xu S, Wang F, Wang K. Evaluation and Optimization Analysis of High-speed Rail Network Structure in Northeast China Under the Background of Northeast Revitalization[J]. Regional Sustainability,2021,2(4):349-362

[10] Zhang P, Zhao Y, Zhu X, et al. Spatial Structure of Urban Agglomeration Under the Impact of High-speed Railway Construction: Based on the Social Network Analysis[J]. Sustainable Cities and Society,2020,62:102 404

[11] 程开明,庄燕杰. 城市体系位序—规模特征的空间计量分析:以中部地区地级以上城市为例[J]. 地理科学,2012,32(8):905-912