分布式混合多态国土空间数据存储与管理技术研究

周海洋
南京市国土资源信息中心　南京　210005

【摘　要】本文针对不同GIS平台对于空间数据重复存储、不兼容的痛点，设计了分布式跨平台空间数据库存储架构，实现一库多平台应用要求，解决了自然资源各部门、各平台、各应用对于混合多态的国土空间数据的统一入库、统一存储、统一应用和统一共享需求，为推动自然资源“两统一”建设提供了一整套的数据解决方案。

【关键词】分布式；混合多态；国土空间数据；自然资源管理

Research on data storage and management technology of distributed hybrid polymorism

Zhou Haiyang
Nanjing Land and Resources Information Center，Nanjing 210005

【Abstract】Aiming at the pain points of repeated storage and incompatibility of spatial data on different GIS platforms，this paper designs a distributed cross-platform spatial database storage architecture to realize the application requirements of one database and multiple platforms，and solves the unified storage，unified storage，unified application and unified sharing of mixed polymorphic land spatial data by various departments，platforms and applications of natural resources，so as to promote the "two unification" of natural resources Construction provides a complete set of data solutions.

【Keywords】cloud-native application architecture；The province coordinates the construction of informatization；Core cloud service development mechanism design；Distributed；Territorial spatial data

1 引言

随着经济和社会活动的高速发展，国土空间数据正日益成为重要的战略资源。然而，这些数据在来源、格式和结构上存在着多样性和异构性的特点，限制了其有效管理和利用的能力，进而带来了一系列现实挑战。首先，存储困难是国土空间数据面临的主要问题之一。由于数据量庞大，传统的存储方法已经无法满足快速增长的需求。因此，学者们致力于开发高效的数据存储技术，包括空间数据压缩算法和分布式存储系统等，以应对数据存储问题。其次，分析困难也是当前国土空间数据面临的挑战之一。由于数据多源异构，数据之间的关联和分析变得复杂而困难。此外，在进行数据处理和分析时，还需要处理数据不准确性、缺失值和异常值等问题，这对分析的可靠性和准确性构成了挑战。因此，学者们正在研究开发各种数据分析和挖掘方法，以从复杂数据中提取有价值的信息。最后，管理困难也是国土空间数据面临的重要问题。数据源众多、格式不一致以及数据交付和使用的复杂性使得数据管理变得复杂而困难。在数据管理方面，学者们致力于开发有效的数据集成和整合方法，以将不同类型的数据整合到一个统一的数据集中，并提出了数据质量评估模型和指标，以确保数据的准确性和可靠性。

近年来，国内外学者在解决国土空间数据多源异构和管理困难的问题上取得了一系列研究进展。在数据集成与整合层面，学者们致力于开发有效的数据集成方法，以实现将不同来源、格式和结构的国土空间数据整合到一个统一的数据集中。例如，刘波等[1]针对地理信息系统多源数据异构冲突导致的整合误差问题，提出基于多源数据集成的地理信息系统数据高效整合方法；孟宏伟等[2]通过对国土规划数据的多源性、异构型、多时空性、多尺度性以及不同坐标系等特点进行阐述和分析，制定基于 GIS 平台的国土规划多源数据集成技术路线。在数据存储与索引层面，为了应对存储困难，学者们提出了各种数据存储和索引技术，以提高存储效率和数据检索速度。例如，戴杨等[3]提出一种分类的数据压缩算法，实现对实时数据库数据的无损和高效压缩；李治君等[4]提出了基于 HBase 的分布式空间数据库存储架构。针对处理多源异构数据时，数据质量和一致性是关键问题，学者们研究了数据清洗、校正和一致性验证等方法，以确保数据的准确性和可靠性。此外，数据质量评估模型和指标的开发也为数据管理提供了重要的支持。对于开放数据与云计算，学者们开始关注如何利用开放数据集和云平台来管理和分析国土空间数据，而开放数据标准的制定和开放数据平台的建设为跨组织和跨领域的数据共享和协作提供了机会和挑战。

面向原规划和原国土数据技术体系不一、数据多源异构向统一管理的难题[5]，以及局内、横向部门、企事业单位和社会公众对于自然资源数据的不同诉求等问题，按照国土空间信息模型数据特点，设计了一种兼顾现状及弹性拓展的分布式混合多态国土空间数据库存储方案。该方案用于存储矢量、栅格、三维、索引、文件、图谱等数据，实现结构化与非结构化、静态与动态、二维与三维等多种类数据的统一存储，满足国土空间信息模型在物理和逻辑上的数据存储与管理需要。本文提出了一种分布式混合多态国土空间数据存储与管理技术，设计了分布式跨平台空间数据库存储架构，实

现一库多平台应用要求，解决了自然资源各部门、各平台、各应用对于混合多态的国土空间数据的统一入库、统一存储、统一应用和统一共享需求，为推动自然资源“两统一”建设提供了一整套的数据解决方案。

2 难点剖析

传统的关系型数据库如 Oracle、MySQL 等，很难满足国土空间三维模型、多媒体数据、栅格数据等非结构化数据的存储与管理，第三方应用的开发必须依赖 GIS 平台的组件，开放性很差。与此同时，商业的数据库软件在分布式计算环境中的部署，还存在成本高、技术复杂和技术封闭等特点，而单主机计算模式在海量数据面前，除了造价昂贵外，在技术上也难于满足数据计算性能指标；同时，随着数据采集方式与手段的快速发展，数据种类、数据内容、数据体量等也呈现高速化的增多趋势，传统单服务器模式的数据存储手段，难以满足多类型、多模态数据存储的要求，这成为极大程度限制数据处理和数据运算的瓶颈。此外，对于国土空间基础信息平台而言，现有的数据存储方式难以适配基于信息七要素与地理七维度的国土空间信息模型，难以支撑全域、全空间、全要素自然资源数据分层分类体系的数据存储。

总体而言，国土空间数据存储与管理主要面临多源数据格式统一难、多维度数据体量大、多部门数据协同更新情况复杂、多需求驱动下数据处理规模扩增、不同的业务数据应用模式不同等难题。

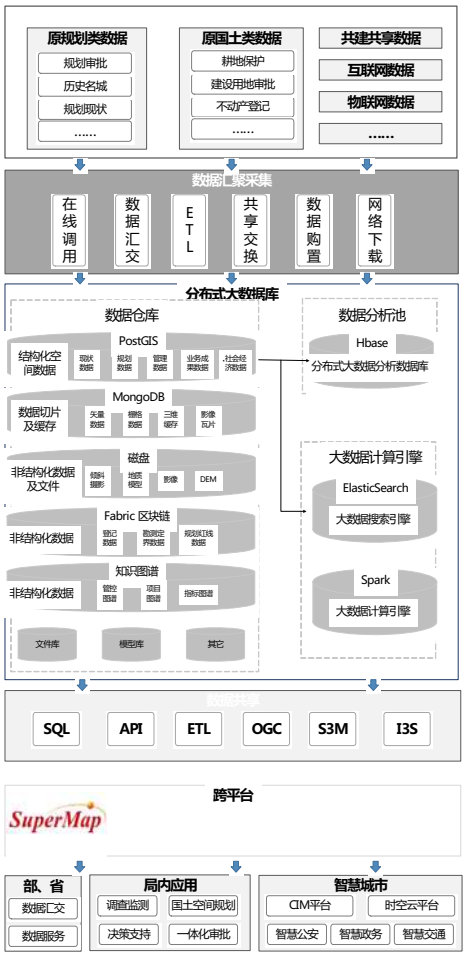

图 1　基础信息平台空间大数据分布式存储方案

3 总体架构

国土资源各类地理信息相关数据种类多种多样，不同种类的数据量大相径庭，基础和专业数据每个专题均超过千万，每个业务管理数据除不动产登记外，数据量较少。不同的业务数据应用模式也不同，包含决策支持、政务服务和信息服务等多种，应用的用户基础有很大区别。因此需要构建不同的数据存储模型，采用合适的存储方式，结合不同的应用需求，保证数据的高效存储与管理。针对不同的数据存储模型，根据数据量和应用需求，采用不同的存储和管理方式。针对非结构化数据，如倾斜摄影、栅格数据、影像、DEM 等，为方便文件的读取、查询、访问，采用 HBase 存储服务；针对结构化数据，一般为关系型数据库中数据，采用基础设施服务中的 PostgreSQL+PostGIS 关系数据库，对于单表超过千万的应用，采用 GreenPlum 分布式存储服务，对海量数据的应用需求，采用 HBase 的方式提高应用效率。其中针对海量智能感知数据，由于其特殊性，采用 ElasticSearch 方式进行管理。利用分布式混合多态国土空间数据存储与管理技术突破空间数据库容量瓶颈，实现空间数据存储管理性能的提升。

4 关键技术

4.1 基于 PostgreSQL+PostGIS 跨平台空间数据存储方法

从目前应用的架构来看，关系型数据库仍然是应用的主要数据存储方式，传统的应用采用 Oracle 或者 SQL Server 等商业数据库存储，由于历史原因等，该类型数据库虽然使用广泛，但是在对空间数据库的支持主要还是由各个 GIS 平台厂商来完成，如 ESRI 的 ArcSDE、SuperMap 的 SDX+ 等，相互之间并不能互操作，第三方应用的开发必须依赖 GIS 平台的组件，开放性很差。与此同时，商业的数据库软件在分布式计算环境中的部署，还存在成本高、技术复杂和技术封闭等特点，而单主机计算模式在海量数据面前，除了造价昂贵外，在技术上也难于满足数据计算性能指标，主机的 Scale-up 模式遇到了瓶颈，SMP（对称多处理）架构难于扩展，并且在 CPU 计算和 IO 吞吐上不能满足海量数据的计算需求。所以如何设计和选择适当的数据库平台，来实现分布式计算要求下的 MPP（海量并行处）是分布式 GIS 服务平台首先要解决的核心问题。

在上述的背景下，本文基于 GreenPlum 设计分布式关系型数据库平台，主要考虑一下几点：

（1）GreenPlum 采用 PostgreSQL 作为底层引擎，良好的兼容了 PostgreSQL 的功能，PostgreSQL 中的功能模块和接口基本上都可以在 GreenPlum 上使用。PostgreSQL 具有非常强大的 SQL 支持能力和非常丰富的统计函数和统计语法支持，除对 ANSI SQL 完全支持外，还支持比如分析函数，还可以用多种语言来写存储过程，对于 Madlib、R 的支持性也很好。

（2）PostgreSQL 的查询优化器非常强大，对于子查询、复制查询如多表关联、外关联等，特别是在关联时对于三大 Join 技术：hash join、merge join、nestloop join 的支持方面功能强大。

（3）扩展性方面，PostgreSQL 也很强大，可以用 Python、C、Perl、TCL、PLSQL 等等语言来扩展功能，另外，开发新的功能模块、新的数据类型、新的索引类型等等非常方便。PG 中 Contrib 目录下的各个第三方模块，在 PostgreSQL 中的 PostGIS 空间数据库、R、Madlib、pgcrypto 各类加密算法、gptext 全文检索都是通过这种方式实现功能扩展的。

（4）PostgreSQL 通过 PostGIS，实现空间数据库引擎，遵循了 OGC 规范，完成了空间查询、运算等多种空间操作，在此基础上扩展了对 3DZ、3DM、4D 坐标的支持，直接通过 SQL 就完成对空间数据的操作。

目前已知的商业 GIS 平台和开源 GIS 平台，均很好的对 PostgreSQL+PostGIS 进行了支持，如 Esri 的 ArcGIS Server、SuperMap iServer 和 GeoServer 等，其开放性和扩展性非常强大。在公有云领域，阿里的时空数据库引擎 Ganos 也是基于 PostgreSQL+PostGIS 的应用。

综上所述，采用 GreenPlum+PostgreSQL+PostGIS 的解决方案已经成为涉及 GIS 应用行业里面的首选架构方案。GreenPlum 基于 PostgreSQL 作为实例（非 Oracle 实例概念，这里指的是一个分布式子库）架构，在 Interconnect 的指挥协调下，能实现数十个甚至数千个 Sub PostgreSQL 数据库实例同时开展并行计算，而且，这些 PostgreSQL 之间采用 share-nothing 无共享架构，从而更将这种并行计算能力发挥到极致。除此之外，GreenPlum 采用两阶段提交和全局事务管理机制来保证集群上分布式事务的一致性，GreenPlum 像 PostgreSQL 一样满足关系型数据库的包括 ACID 在内的所有特征。整个分布式关系型数据库的架构如图所示。

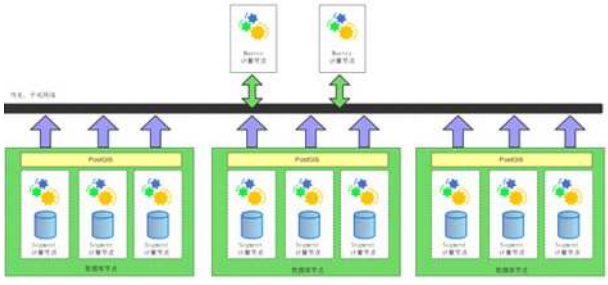

图 2 GreenPlum+PostgreSQL+PostGIS 数据存储架构

4.2 基于 MongoDB 二三维地图高性能缓存存储技术

MongoDB 作为 NoSQL 中面向文档的分布式存储数据库，具备模式自由、灵活性高的特点，支持完全索引以及海量数据高效存储，具备云计算级别的扩展性、高可用性、高并发读写性能、经济性等独到的优势，为非结构化数据的高效组织与存储提出了一种新的解决思路。采用 MongoDB 方式可以更有效地管理非结构化文件数据，其支持文件压缩、加密等功能。

针对海量基础时空数据\的存储、查询、检索和并行处理等问题，利用 MongoDB 数据库文档数据模型、无模式特性以及与云计算平台的交互能力，提出适用于矢量空间数据的云存储与处理方法，通过 MongoDB 部署，实现二维瓦片和三维模型缓存存储，同时满足用户对异构矢量空间数据存储与网络服务的高性能需求。

MongoDB 二三维地图高性能缓存存储主要用途包括：二维瓦片地图存储、三维模型缓存存储。其中，二维瓦片地图存储主要实现创建缓存，在 Supermap iDesktop 中创建 MongoDB 缓存，修改 iDesktop 桌面的[生成地图缓存]窗口[输出设置]；文件结构，二维缓存保存至 MongoDB，对应用户下有多张表，主要是源数据和不同比例尺的瓦片数据。源数据表中主要存储 sci 相关信息。瓦片数据表依据指定的行列号存储对应比例尺下的图片二进制信息。三维模型缓存存储可将 OSGB 数据保存到 MongoDB 中，SuperMap iDesktop9D 版本提供将 OSGB 数据保存至 MongoDB 的功能。

4.3 基于 HDFS/HBase/Spark 时空大数据存储与分析技术

基于空间模型的大数据分析技术，形成二维、三维、静态和实时动态可视化展示应用成果，辅助宏观决策。其中，多源数据的接入可支持 SuperMap、ArcGIS 多种数据源；而分布式高效计算则支持亿万级矢量数据的分布式空间计算，相对于传统计算模式存储效率具有显著的提升。

HBase 是高可靠性、高性能、面向列、可伸缩的分布式非关系型存储系统，区别于传统的关系型数据，是面向列存储模型的分布式数据存储系统，支持结构化及非结构化数据，利用 HDFS 作为其文件系统，可以支持海量空间数据存储[6]。HBase 与传统关系型数据库最大的区别是表的行列设计上，以列簇的形式分组列，列簇在物理上都是独立文件及配置存储；HBase 的行为逻辑上的行是依照物理上的列簇分别存取的，RowKey 则作为连接各个列簇行的关键；同时，HBase 中的数据有版本的概念，数据修改都会保存对应时间戳版本，版本数量可以由表定义设定。

采用 HBase 作为空间数据库的存储方式，和传统的关系型数据库空间数据库引擎一样，需要实现四个方面的内容，包括空间对象描述、空间对象序列化、空间对象索引和空间对象查询。

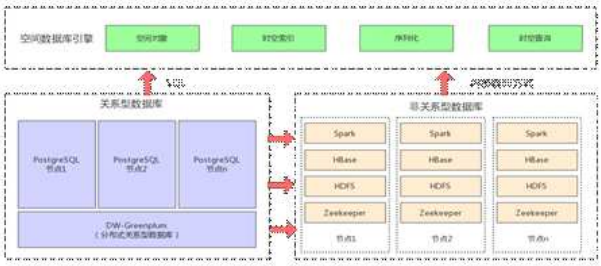

图 3　HBase 空间数据库引擎构成

通过构建分布式混合多态国土空间数据存储与管理技术研究，支撑南京市规划和自然资源数据管理，实现了数据的有效存储和更新。运用 PostgreSQL 库存储技术、HBase 存储技术、大数据分析技术等，完成建设 PostgreSQL 数据库实例 3 个、HBase 数据库集群（实例）1 个，分布式大数据节点 4 个、HDFS 数据库实例 4 个、ElasticSearch 数据库实例 1 个，满足多类型、多模态数据的存储与管理。

5 建设效益

表 1　平台数据存储方式

序号	详细分类	存储方式	备注
1	结构化矢量数据（小于 1000 万）	PostgreSQL+PostGIS	–
2	结构化矢量数据（大于 1000 万，小于 5000 万）	PostgreSQL+PostGIS+GreenPlum	–
3	结构化矢量数据（大于 5000 万）	HBase	–
4	智能感知等物联网数据	ElasticSearch	–
5	多时相影像数据	HBase	–
6	三维数据切片及缓存	MongoDB	–
7	敏感数据	Fabric	不动产登记等
8	文件型数据	HDFS	
9	图数据	HugeGraph	

6 总结与展望

面向原规划和原国土数据技术体系不一、数据多源异构向统一管理的难题，以及局内、横向部门、企事业单位和社会公众对于自然资源数据的不同诉求等问题，按照国土空间信息模型数据特点，设计了一种兼顾现状及弹性拓展的分布式混合多态国土空间数据库存储方案。该方案用于存储矢量、栅格、三维、索引、文件、图谱等数据，实现结构化与非结构化、静态与动态、二维与三维等多种类数据的统一存储，满足国土空间信息模型在物理和逻辑上的数据存储与管理需要。分布式混合多态国土空间数据存储与管理技术研究，有助于进一步推进资源整合、系统互联互通和数据共享，节约信息化建设成本，能更好地支撑自然资源数字化政府建设，提高行政管理效能。截至目前，本研究的成果已在南京市国土空间基础信息平台等多个信息化项目中进行应用并取得实际效果。下一步，将结合南京实际，加强创新技术应用，深入探索云原生应用架构模式在信息化建设中的更广泛应用，推进市自然资源信息化建设水平的提高。

参考文献

[1]刘波.基于多源数据集成的地理信息系统数据高效整合研究[J].经纬天地，2021，No.202（05）：93-96.
[2]孟宏伟.基于 GIS 平台的国土规划多源数据集成应用实例[J].测绘与空间地理信息，2018，41（01）：182-184.
[3]戴杨，陈芳.实时数据库中数据的分类压缩算法[J].计算机与现代化，2016，No.250（06）：123-126.
[4]李治君，周俊杰，范延平等.国家级国土空间基础信息平台分布式数据库设计与实现[J].自然资源信息化，2022，No.131（05）：80-85.
[5]陈军，武昊，张继贤等.自然资源调查监测技术体系构建的方向与任务[J].地理学报，2022，77（05）：1041-1055.
[6]龚敏霞，吉波，徐年峰，胡岭.自然资源分布式 GIS 服务平台研究[J].现代测绘，2020，43（05）：21-24.