DOI:10.11737/j.issn.1003-0239.2024.06.006

针对台风风暴潮业务的多源多维数据快速提取技术

刘思晗[1,2]，蔡文博[1,2*]，李飞[1]，王凤菊[1]，叶文琦[3]

（1.国家海洋环境预报中心，北京 100081；2.国家海洋环境预报中心 自然资源部海洋灾害预报技术重点实验室，北京100081；3.国际商业机器（中国）有限公司，北京 100027）

摘 要： 分析当前海洋大数据处理及提取存在的信息化问题，以台风风暴潮预报系统为例阐述了多源异构海洋环境数据的持久化保存以及快速提取关键技术。针对台风风暴潮预报业务数据读取服务场景进行深入分析，并针对该业务的海洋环境数据存储和提取提出了基于OldSQL+NoSQL+分布式文件系统的解决方案。该方案经过实际场景验证，可以有效解决台风风暴潮数值预报业务化系统中海量多源异构数据的存储和提取问题，显著提升数据提取性能，可扩展性较强。

关键词： 台风风暴潮；数据管理；海洋环境数据；关系型数据库

中图分类号：P731.23 **文献标识码：**A **文章编号：**1003-0239（2024）06-0053-09

0 引言

近年来充分利用智能现代化信息技术构建的各类海洋预报业务系统被广泛使用[1]。海洋大数据是观测或计算得到的不同时空尺度的海洋信息，是辅助了解海洋状态、发现海洋过程及规律、解决海洋系统所面临挑战的基础，其核心能力是预测未来一段时间的海洋环境、气候及资源的时空变换[2]。海洋大数据存在体量大、维度高、动态性、时空相关性、多尺度以及异构性等特征[3]。

海洋环境预警报业务会不断产生大量庞杂的预报产品（数据），且生成的数据多以多维异构空间数据为主。为了满足业务主体对不同预报数据集的集成、统一和科学保存的需求，亟需提升预报数据的易用性和可解释性，以解决各类预报业务中操作繁琐、数据解读困难等问题。

海洋环境要素数据由于来源不同、时间跨度较长、数据种类繁多且多数以文件形式进行存储，因此存在数据格式异构、有效存储和高效检索提取困难的问题[4]。

构建台风风暴潮数值预报业务化系统需要面对如何高效持久地保存海量多源异构数据、构建合理的存储系统以及有效地将预报数据与实际业务系统相关联等挑战[5]。本文针对台风风暴潮预报过程中产生的多元化数据，通过分解预报流程中产生的各类数据，探索适用于台风风暴潮业务化系统中模型场、单点时序数据以及相关联的气象数据信息的持久化保存及提取等整套解决方案。

1 多源多维数据持久化保存及提取

多源数据（Multisource Data）是指来自于不同来源或系统的数据集合，或来自不同数据库的结构化和非结构化的数据集。多源数据可能具有不同的格式、结构和语义，需要进行整合和处理才能得到有意义的信息。多维数据是由多个维度组成的数据集合，每个维度代表数据的一个属性或特征，而数据集合则由这些维度交叉组成。

海洋环境数据多为多源多维数据。目前主流的基于数据库的存储技术有基于事务处理的OldSQL、适用于数据分析应用的NewSQL以及适用于互联网领域的NoSQL[6]。针对此类数据结构，宋

收稿日期： 2024-02-18。
基金项目： 国家自然科学基金（42076214）。
作者简介： 刘思晗（1987-），男，工程师，硕士，主要从事海洋数据挖掘及WebGIS系统研发与研究工作。E-mail: evaseemefly@126.com
*通信作者：** 蔡文博（1983-），男，高级工程师，硕士，主要从事海洋预警报大数据管理与应用工作。E-mail: 19345288@qq.com

晓等[7]提出了针对非结构化数据、时间序列数据以及空间数据等多种类型的海洋数据，采用“OldSQL+NewSQL+NoSQL”混搭模式的数据库存储方式，以及HDFS和MapReduce基于Hadoop大数据系统架构实现对海洋数据的存储与管理。针对海洋环境数据多源多维的特点，目前的存储方式一种是提取数据并存储在关系型数据库中，另一种是基于文件系统的通用数据模型建立数据存储与管理服务，提供一致性的数据访问协议。

本文设计的存储系统（简称本设计系统）针对海洋环境大数据多源多维特性，以结构化数据文件（如NetCDF等）为基础，提取文件中的基础信息和特征数据并存储至关系型数据库中（如MySQL），再通过ORM模型建立数据（实体）与数据库（表）的映射关系，将结构化数据以垂直分片等方式进行存储[8]后用于读取结构化数据文件，同时结合联合索引、查询优化、缓存（如Redis）和并行处理（如Dask）等技术。对部分文档型数据文件预先加载提取，并以模型对象映射的方式写入数据库，以便在提取时直接根据特征关系从数据库提取。

本设计系统最终采用OldSQL+NoSQL+分布式文件系统的混合存储方式，以实现对多源多维结构化数据的快速存储与提取。这种方式可以提高数据访问效率，并支持更灵活高效的数据分析和处理。

2 存储方案设计

台风风暴潮预报数据主要包括风暴潮增水场、提取后的单点潮位站和台风路径气象数据，这些数据具有物理特征。3类数据分别以关系型数据库（OldSQL）和分布式文件系统进行存储，而对于访问频次较高的少量数据则存储至缓存中（NoSQL）。为了便于相应的服务系统进行提取，存储的文件通过关系型数据库建立索引关系。

2.1 模型场数据的存储

基于不同种类可将台风风暴潮增水场预报产品分为最大增水场预报、逐时增水场预报与概率增水场预报。以最大增水场产品为例（见图1），它是一个多维数据，能够以结构化的NetCDF文件进行存储，该数据包含经纬度（lat，lon）、时间（time）3个变量（多维度），其中经纬度（lat，lon）以矩阵的方式进行表述。针对模型场数据体量大、维度高、动态性、时空相关性、多尺度以及异构性等特点，存储设计需要考虑多方面因素。针对风暴潮预报业务，存储设计时需要权衡存储空间与数据提取效率的动态平衡并加入冗余设计。对于结构化数据，一般采用切分降维，并结合关系型数据库建立联系，以实现存储设计。

本设计系统基于NetCDF文件为主要载体对预报场进行持久化保存，并配合数据库实现数据的快速读取。具体步骤包括：数据预处理、构建数据索引、基于geotiff格式的数据切片与转存、混合查询、基于结构化数据的按需加载（即懒加载）、并行处理、缓存机制。具体流程如下：

S1：将每个时次发布的各类增水场（最大增水场、概率增水场、逐时增水场）的NetCDF文件数据进行解析、提取特征数据与基础信息（如NetCDF文件中的变量、维度等信息）映射到数据库表的相应字段中并存储到数据库中。

S2：在关系型数据库中为NetCDF文件中的不同物理变量与时间维度创建索引以提升数据检索的效率。根据数据的特点和查询需求创建适当的索引，例如基于时间、空间或其他关键属性信息。

S3：将各类增水场数据按照所提取的特征（增水概率、时间等）进行转存与切片，将数据按照特定要素或特征进行切分并转换为geotiff格式。以概率增水场为例，按照概率维度进行切分（即区分不同增水概率）后转存为geotiff格式文件，并在关系型数据库中记录大于每个概率对应的geotiff文件的特征信息。

S4：根据预报的时间范围与变量种类，通过数据库查询获取指定的geotiff格式文件与结构化数据特征信息，读取结构化数据并按需提取（即懒加载）。

S5：通过混合查询确认待加载的变量和维度，将结构化数据按需提取至内存中（即实现按需加载）。

S6：利用并行计算和多线程技术，同时从数据库和NetCDF文件中读取大量数据，以提升数据读取和聚合速度（使用Dask与xarray库实现并行与懒加载处理）。

S7:在设计的“管道”(Queue——先进先出的管道设计模式)中,将已读取的关键数据缓存至内存中,以实现对高频读取的快速访问。

对于分辨率较高的NetCDF数据,可以适当地降低其分辨率并存储为geotiff格式。通常可采取重采样处理,将分辨率按需降低到合理水平。重采样算法包括立方卷积(Trilinear Interpolation)、最临近差值(Nearest-neighbor Inter- polation)以及双线性插值(Bilinear Interpolation)等。

2.2　单点潮位数据的存储

潮位预报结果以结构化数据形式呈现,使用约定的分隔符进行分割,并存储了计算后对该站点影响的全部时刻的时序数据。单点潮位数据存储在关系型数据库中,其建立在关系模型基础上,通过一对一或者多对多等形式建立关系,支持事务处理和数据的持久化储存,通常遵循三范式的设计思想。

对于信息多元且结构复杂的海洋时空数据,通常会建立数据索引(特征/时间/空间)。常见的索引技术包括二叉树类型索引以及基于B树索引[9]。本设计系统针对台风风暴潮不同要素的预报数据,按照要素信息进行存储并建立联合索引。在使用的关系型数据库中,采用B+树索引建立联合索引,B+树在查找效率和查找范围方面都具有出色的性能。

将海洋站基本信息作为静态信息进行存储,将各个单点潮位站数据和预报数据分表存储。

在大型数据库中,设置外键会增加数据库查询和更新操作的开销,也会限制数据库的操作,还会

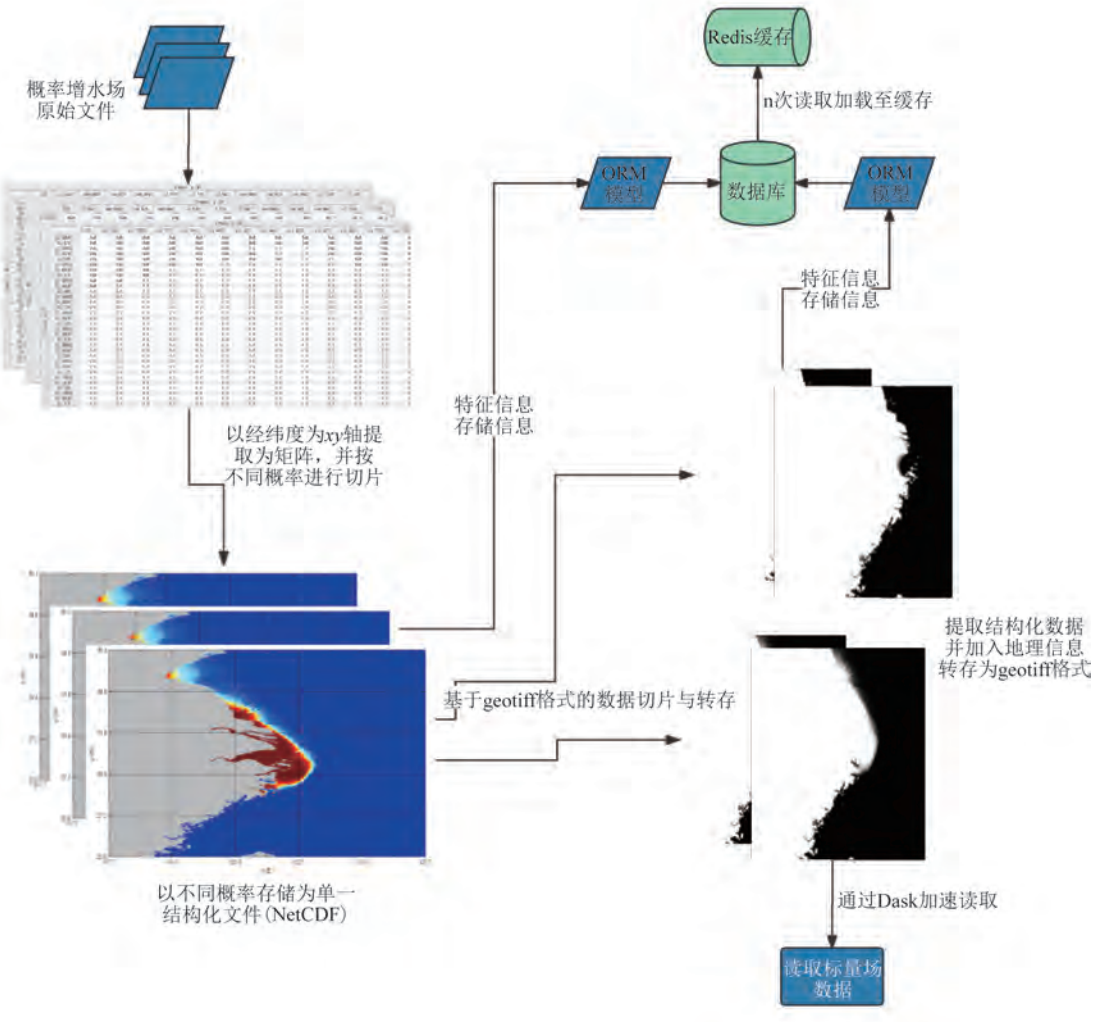


图1　概率增水场数据快速提取方案

Fig.1　Fast extraction solution for storm surge probability data

引发复杂业务逻辑中数据不一致的风险[10]。因此本设计系统在设计时尽量减少设置外键,采用虚拟外键和虚拟联合外键的形式关联逻辑上的主从表。由于本设计系统数据量较大,且更新或新增操作只在指定时间或低频业务中进行,因此建立了联合索引。以海洋站存储业务为例,海洋站数据的数据库表及主要功能详见表1。以潮位预报业务为例,在警戒潮位表(见表1中T3)中,通过站点编号(station_code)与预报时间戳(forecast_ts)便可确定指定站点的警戒潮位(4种对应警戒潮位及对应水尺高度换算关系等);在风暴增水表(T2)中,通过站点编号(station_code)、集合路径ID(group_id)、作业创建时间戳(timestamp)与对应的预报时间戳(forecast_ts)即可确定指定站点指定预报时间的增水详情;在统计分位数表(T4)中,通过站点编号(station_code)、作业创建时间戳(timestamp)与对应的预报时间戳(forecast_ts)即可确定指定作业创建的各路径在指定预报时刻的数据分布情况。

相关数据库表及主要字段见图2。从数据库关系可见,单点潮位预报业务包含了台风基础信息、台风预报信息、集合预报路径信息、海洋站基础信息、天文潮基础信息、警戒潮位基础信息以及不同集合预报路径对应的站点增水信息。这其中,台风基础信息以台风ID为虚拟外键,台风预报信息与集合预报路径信息以group_id为虚拟外键,建立与被影响的海洋站增水的关系;海洋站增水以站点编码(station_code)为虚拟外键,与海洋站基础信息、天文潮信息及警戒潮位信息表建立联系。

本设计系统在读取和存储关系型数据库时采用了对象关系映射(Object-Relational Mapping,ORM)模式,即在关系型数据库与业务实体对象之间建立映射[11]。ORM可以将数据库表与面向对象语言中的类进行映射,开发人员可以使用面向对象的方式来操作数据库。本设计系统台风集合路径及增水业务涉及的主要ORM模型及继承关系见图3。

2.3 台风气象及路径信息数据存储

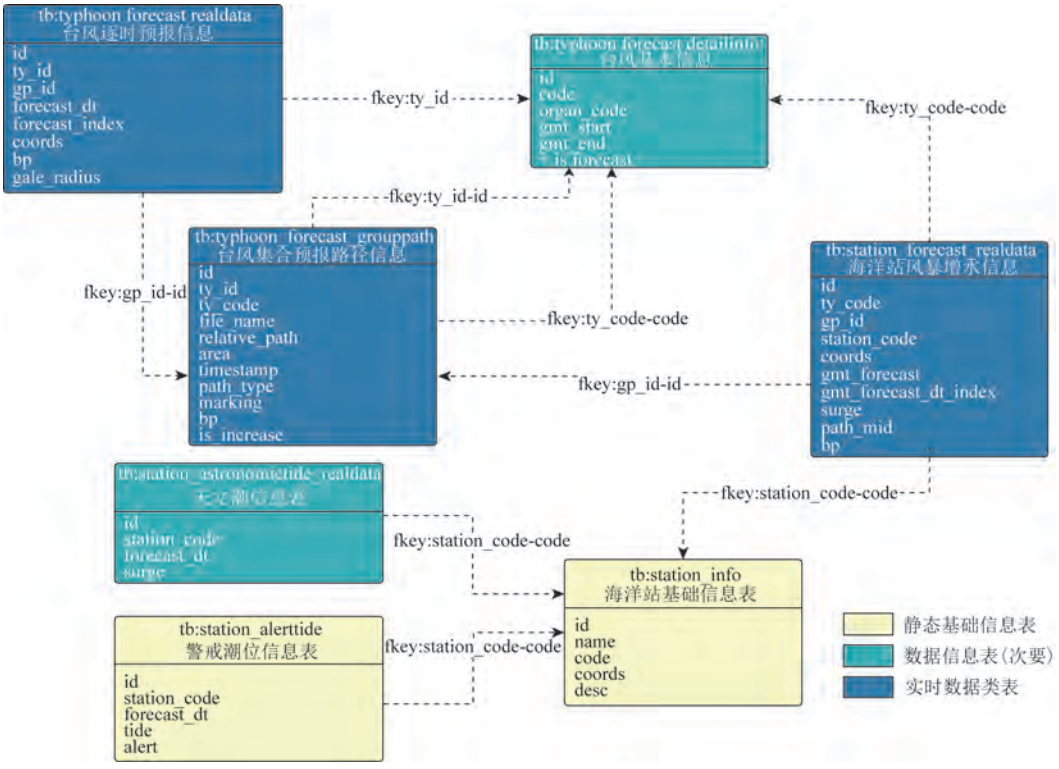
与台风相关联的气象信息数据常存储在关系型数据库中[12-13],这类数据主要包含台风路径信息以及必要的台风气象信息,例如台风中心经纬度、气压、大风半径等;台风集合预报结果包含145个预报成员的路径偏移位移标识、气压、气压增减量等信息。对于台风路径的存储,由于同一个台风可能由于预报发布时间不同而产生多个预报结果(预报

表1 站点增水预报业务数据库表及功能描述

Tab.1 Database table and function description of station water surge forecasts

编号	表名称	功能描述
T1	海洋站基础信息表 (station_info)	存储海洋站的基础信息:海洋站名称、编码、经纬度信息、描述信息以及所属父级机构代码等。
T2	海洋站风暴增水表* (station_forecast_realdata_xxxx)	由于某次过程不同时间的台风中心预报路径会有所偏差,导致集合预报的结果随着时间推移有所变化。对于同一个台风会存在多个时刻的预报路径信息,所以将台风编号与时间戳(获取的中央气象台台风中心路径时间)作为联合索引。海洋站风暴增水表主要用来存储不同时刻的台风索引信息,包括集合预报路径的ID、海洋站编码、预报时刻以及对应风暴增水值、集合预报台风的气压增减值等信息。
T3	海洋站警戒潮位表 (station_stationalerttide_data)	主要存储每个海洋站不同警戒级别对应的风暴增水值。数据为静态数据。
T4	海洋站统计分位数表 (station_qunatile_realdata)	主要存储潮位站由台风中心路径计算得到的各条集合预报路径的统计结果。主要包含四分之一百分位数、中位数、四分之三百分位数与全部集合预报最大值与最小值范围。
T5	海洋站天文潮位表 (station_astornomictide_realdata)	存储海洋站的天文潮位表,静态数据,每年更新一次,主要包含海洋站编码、预报时刻、对应的天文潮高等信息。

注:*按照台风编号进行分表处理



*虚线为虚拟外键,实际关联表未建立实体外键

图2　站点增水业务相关数据库表关系

Fig.2　Database table relationships for station surge forecasts

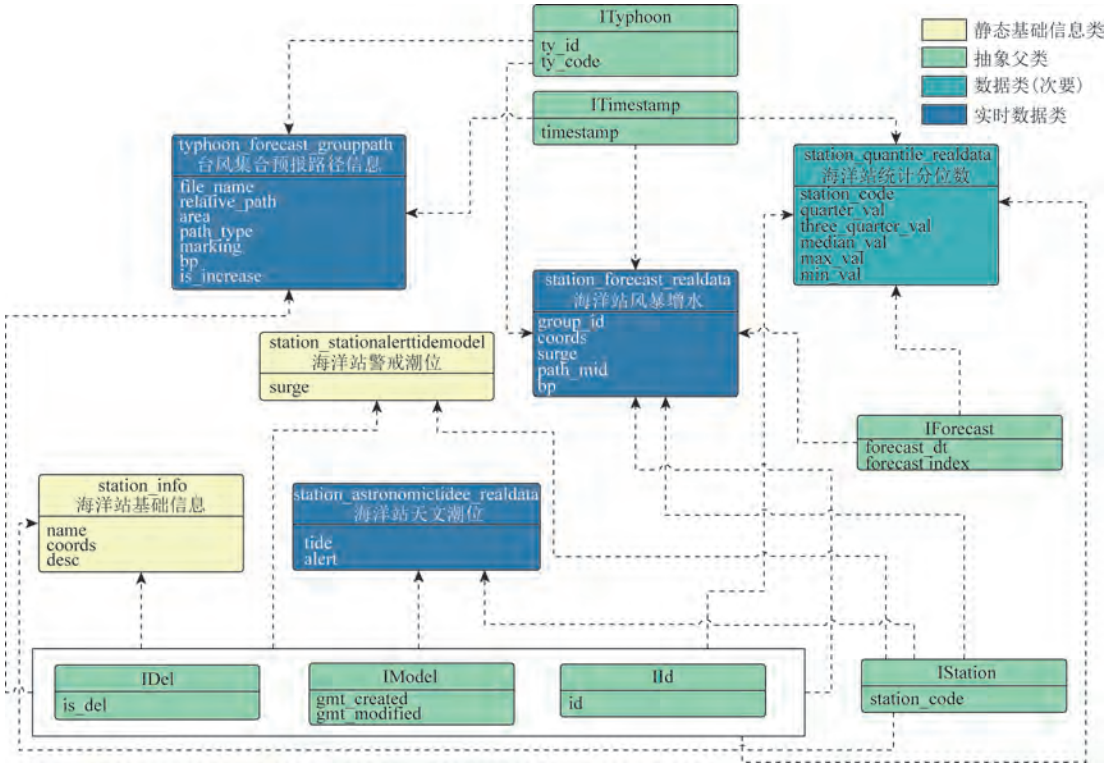


图3　ORM模型及继承关系

Fig.3　ORM model and inheritance relationship

路径及相关气象信息),所以区分路径的标识信息应为台风编号+获取时的时间戳+混淆数字(加入混淆随机数)。

台风基础信息表(typhoon_detail)的功能为存储发布时刻的台风基础信息,包括台风ID*(*表示主键,下同)、台风编号、发布机构代码、预报来源、预报实况、发布时间戳以及其他信息。台风集合预报路径群组表(typhoon_group_path)的功能为存储基于指定台风的集合预报路径信息,包括集合预报路径ID*、台风ID#(#表示引用其他表的主键,但未设置为外键,下同)、区域、相对路径、文件名、预报时间戳、路径类型(中间、偏左、偏右)、路径偏移量、路径是否提前,气压以及其他信息。台风预报数据(typhoon_forecast_data)的功能为基于指定时刻发布台风预报及实况路径等信息,包括台风ID#、集合预报路径ID#、预报时间、预报时间对应的当前发布总预报时次的索引编号、经纬度、大风半径、时间戳以及其他信息。

台风基本信息表中的一条数据与台风集合预报路径群组存在一对多的关系;台风集合预报路径群组表中的一条数据分别与台风预报数据表和海洋站风暴增水数据表以及其他单点数据表存在一对多的关系。台风气象及路径信息关系见图4。

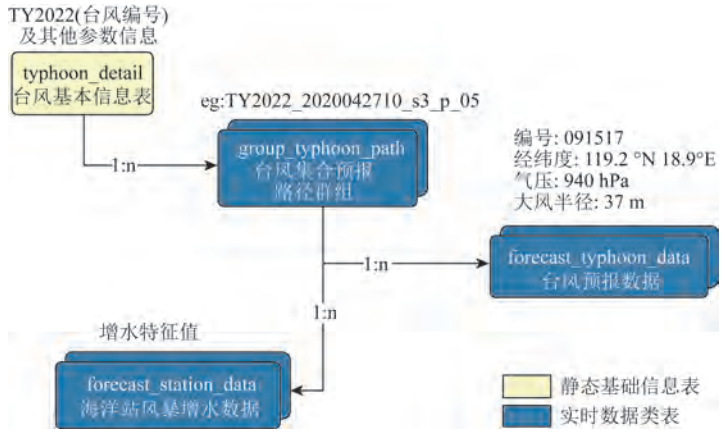


图4 关键表之间的对应关系简图

Fig.4 Diagram of the relationship between key tables

3 数据查询性能测试

以2212号台风"Muifa"(梅花)过程为例,共发布风暴潮警报5期,使用本设计系统提交9次计算作业,产生文本文件及结构化文件3 420个,数据大小为2.42 Gb。该例涉及的主要统计信息为海洋站风暴增水表、台风集合预报路径群组表与海洋站天文潮表。各库数据量见表2。部分数据服务查询效率对比图见图5,详细查询响应时间见表3。

经过分析发现,通过分表和基于B+树建立联合索引的方式可以提高海洋时空数据的存取效率,特别是针对信息多源多维、结构复杂的台风风暴潮大规模数据(一次台风过程影响站点的增水对应表+天文潮表+其余关联表总数据量近1 000万行)。B+树索引可以进行大量存储并减少磁盘I/O操作的次数,按需建立联合索引并遵循左侧原则可以显著提升查询性能。本设计系统目前只应用于近3年的台风预报业务,随着后续数据量的持续增加,需要加入主从复制、读写分离等设计方案以保证查询效率。

针对前台页面提取增水场等矩阵并渲染成栅格图层的业务逻辑,一般采用远程加载geotiff格式文件的形式进行,减少直接将矩阵存储至关系型数据库并读取而增加的I/O操作,通过在局域网内远程加载geotiff格式文件可保持较高的处理效率。

表 2　各库表数据量

Tab.2　Data mass of various databases

编号	表名	统计条件	结果/行
S1	T1	2212号台风下各海洋站不同预报时刻各个集合路径的全部预报数据	6 645 785
S2		2212号台风作业创建时次	9
S3		所有站点一个作业时次全部预报路径数据	793 875
S4	T3	天文潮表总数据量，起止时间为 2019年12月31日16时（世界时）—2023年12月31日15时	3 556 800
S5	T3	天文潮表2022年总数据量	1 103 736
S6	T2	所有台风集合预报路径数据	21 500
S7		2212号台风全部提交作业的集合预报路径	1 305
S8		2022号台风某一作业时次创建的集合预报路径数量	145

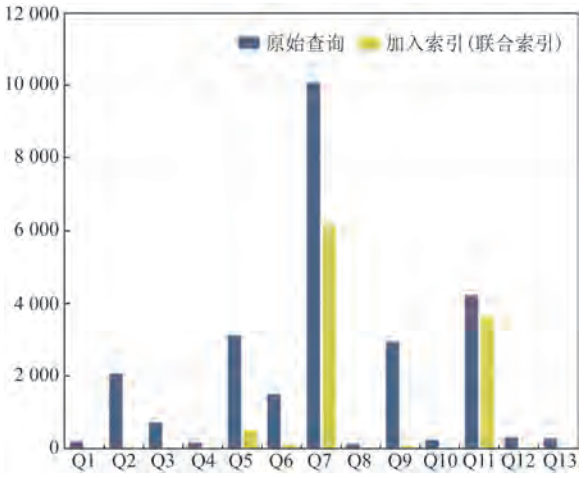


图5　部分请求查询整体响应时间

Fig.5　Overall response time for partial request queries

表3　部分请求查询整体响应时间

Tab.3　Overall response time for partial request queries

编号	服务器响应时间	功能描述
Q1	190.49 ms	获取指定年份(year=2022)已创建的台风编号集合
Q2	2.07 s	获取指定台风编号(ty_code)的所有作业集合
Q3	718.39 ms	获取指定台风ID(ty_id—指定作业创建的台风路径)的集合路径集合
Q4	166.35 ms	根据台风编号(ty_code)以及创建时间戳(ty_timestamp)获取指定作业对应的最大增水场tif url地址
Q5	3.12 s	根据台风编号(ty_code)与创建时间戳(ty_timestamp)获取该过程的站点极值
Q5-P*	510.6 ms	
Q6	1.5 s	根据台风编号(ty_code)以及创建时间戳(ty_timestamp)获取指定站点的天文潮位
Q6-P*	108 ms	

表3（续）

Tab.3（Continued）

编号	服务器响应时间	功能描述
Q7	10.07 s	根据台风编号(ty_code)与创建时间戳(ty_timestamp)加载单站(station_code)预报的中间路径增水与集合路径预报的极值(max/min)
Q7-P*	6.12 s	
Q8	127.09 ms	获取单站(station_code)警戒潮位
Q9	2.94 s	获取指定台风集合预报路径(gp_id)、预报时间(forecast_dt)以及创建时间戳(ty_timestamp)对应的全部站点的增水集合
Q9-P*	73 ms	
Q10	238.64 ms	根据台风编号(ty_code)与创建时间戳(ty_timestamp)获取对应作业概率增水场(coverage_type)tif url 地址
Q11	4.22 s	根据台风编号(ty_code)与创建时间戳(ty_timestamp)获取对应作业指定单站(station_code)所有集合预报的增水
Q11-P*	3.63 s	
Q12	312 ms	远程加载指定url对应的最大增水场tif文件,文件大小为2.9 MB
Q13	289 ms	远程加载指定url对应的概率增水场(大于1.0 m) tif文件,文件大小为2.9 MB

4 结论

本文结合台风风暴潮预报业务,针对海洋环境的大规模数据提出了数据持久化保存与快速提取方案。针对台风风暴潮预警报业务中涉及的模型场数据、单点潮位数据以及台风气象及路径信息数据,通过OldSQL+NoSQL+分布式文件系统构建的存储方案,有效满足了风暴潮预报业务系统的数据查询、检索与分析需要。

本文对海洋环境大规模数据的管理及存储提取进行了深入探索与研究,并验证了该存储方案在海洋环境大规模数据存储领域的高效性。这一技术路线可为其他海洋预报系统中大规模数据的管理与存储提供借鉴与参考。

参考文献:

[1] 李雪丁, 曾银东, 陈金瑞, 等. 福建省智能网格海洋预报业务系统实现与应用[J]. 海洋预报, 2021, 38(1): 10-17.
LI X D, ZENG Y D, CHEN J R, et al. Establishment and application of an intelligent grid operational marine forecasting in Fujian province[J]. Marine Forecasts, 2021, 38(1): 10-17.

[2] 刘帅, 陈戈, 刘颖洁, 等. 海洋大数据应用技术分析与趋势研究[J]. 中国海洋大学学报, 2020, 50(1): 154-164.
LIU S, CHEN G, LIU Y J, et al. Research and analysis on marine big data applied technology[J]. Periodical of Ocean University of China, 2020, 50(1): 154-164.

[3] 于婷, 董明媚, 殷悦. 海洋科学大数据共享与价值研究[J]. 海洋信息, 2021, 36(3): 31-42.
YU T, DONG M M, YIN Y. On the marine big data sharing and value mining[J]. Marine Information, 2021, 36(3): 31-42.

[4] 沈飞飞, 郭忠文, 胡克勇. 一种海洋环境数据存储模型设计与应用[J]. 中国海洋大学学报, 2015, 45(6): 122-127.
SHEN F F, GUO Z W, HU K Y. Design and application of a marine environmental data storage model[J]. Periodical of Ocean University of China, 2015, 45(6): 122-127.

[5] 刘贤三, 张新, 董文, 等. 风暴潮灾害预报的数据组织技术[J]. 自然灾害学报, 2010, 19(2): 136-139.
LIU X S, ZHANG X, DONG W, et al. Data organization technology for prediction of storm surge disaster[J]. Journal of Natural Disasters, 2010, 19(2): 136-139.

[6] 韦广昊, 韩春花, 田先德, 等. 海洋大数据管理技术: 架构、平台与存储策略[J]. 海洋信息技术与应用, 2021, 36(4): 46-54.
WEI G H, HAN C H, TIAN X D, et al. Marine big data management technologies: architecture, platform and storage strategy[J]. Journal of Marine Information Technology and Application, 2021, 36(4): 46-54.

[7] 宋晓, 梁建峰, 李维禄, 等. 基于多架构混搭模式的极地海洋数据库建模技术研究[J]. 极地研究, 2018, 30(4): 411-418.
SONG X, LIANG J F, LI W L, et al. Developing databases with multiple architectures to support polar marine data[J]. Chinese Journal of Polar Research, 2018, 30(4): 411-418.

[8] 张玉娟, 史绍雨, 孙晶, 等. 基于分布式数据库的海洋动力环境数据云存储[J]. 海洋预报, 2017, 34(2): 72-79.
ZHANG Y J, SHI S Y, SUN J, et al. Cloud storage of ocean dynamics environment data[J]. Marine Forecasts, 2017, 34(2): 72-79.

[9] 谭凯中, 秦勃, 何亚文. 面向过程的海洋时空数据分布式存储与并行检索[J]. 中国海洋大学学报, 2021, 51(11): 94-101.

TAN K Z, QIN B, HE Y W. Process-oriented distributed storage and retrieval of ocean spatiotemporal data[J]. Periodical of Ocean University of China, 2021, 51(11): 94-101.

[10] SCHWARTZ B, ZAITSEV P, TKACHENKO V. High performance MySQL[M]. 3rd ed. Sebastopol: O'Reilly Media, Inc., 2012: 141-152.

[11] HALPIN T, BLOESCH A. Data modeling in UML and ORM: a comparison[J]. Journal of Database Management, 1999, 10(4): 4-13.

[12] 陈钻, 李海胜. 新型台风海洋网络气象信息系统的设计与实现[J]. 应用气象学报, 2012, 23(2): 245-250.
CHEN Z, LI H S. The design of typhoon and marine meteorological information system with its implementation[J]. Journal of Applied Meteorological Science, 2012, 23(2): 245-250.

[13] 王舒, 陈京华, 吉玮, 等. 气象地理信息系统设计与实现[J]. 地理空间信息, 2020, 18(2): 5-8.
WANG S, CHEN J H, JI W, et al. Design and implementation of meteorological geographical information system[J]. Geospatial Information, 2020, 18(2): 5-8.

Rapid data extraction technology for multi-source multi-dimensional data in typhoon storm surge forecasts

LIU Sihan[1,2], CAI Wenbo[1,2*], LI Fei[1], WANG Fengju[1], YE Wenqi[3]

(1. National Marine Environmental Forecasting Center, Beijing 100081, China; 2. Key Laboratory of Marine Hazards Forecasting, National Marine Environmental Forecasting Center, Ministry of Natural Resources, Beijing 100081, China; 3. International Business Machines Corporation, Beijing 100027, China)

Abstract: This paper analyzes the informatization management in the marine big data age, and describes the key technologies for persistent storage and rapid extraction of multi-source heterogeneous marine environmental data in typhoon storm surge forecasts. This study proposes an OldSQL+NoSQL+distributed file system based solution for marine environmental data storage and extraction in typhoon storm surge forecasts. The solution has been verified in actual scenarios and has been proven to be efficient and scalable in solving the massive multi-source heterogeneous data storage and extraction in the typhoon storm surge forecasts.

Key words: typhoon storm surge; data management; marine environmental data; relational database