

突发水污染事件中遥感瓦片大数据存储系统^①

李维乾¹, 解建仓², 李建勋³, 李莉¹

¹(西安工程大学 计算机科学学院, 西安 710048)

²(西安理工大学 水资源研究所, 西安 710048)

³(西安理工大学 经济与管理学院, 西安 710048)

摘要: 突发水污染事件频发, 利用 GIS 平台、遥感技术及水质模型对突发水污染事件进行模拟研究逐渐成为一种发展趋势. 针对目前海量、异构、动态遥感数据难以快速处置和高效存储的问题, 基于 HBase 设计和实现了突发水污染事件的遥感瓦片大数据存储系统. 该系统以数字地球平台为支撑, 结合遥感瓦片金字塔模型和 MapReduce 框架, 考虑实时、动态观察事件发展态势引起的瓦片数据频繁加载特点, 对遥感瓦片数据的线性二叉树编码索引进行了改进, 将其同水污染数据一并存储到 HBase 数据库中, 并通过增加缓存区来提升遥感瓦片数据的访问效率. 将此系统应用于渭河陕西段突发水污染事件中, 实验结果表明: 改进后的索引可在 200 个毫秒时间内提取到水污染区域的遥感数据, 引入的缓存机制使得水污染区域遥感数据的加载时间节省了近 2/3. 因此, 该存储系统访问效率高, 能够满足突发水污染事件快速应急模拟的需求.

关键词: 突发水污染事件; 遥感瓦片数据; 大数据; MapReduce; HBase

Storage System on the Remote Sensing Tile Data Based on Big Data Technology in Sudden Water Pollution Incident

LI Wei-Qian¹, XIE Jian-Cang², LI Jian-Xun³, LI Li¹

¹(College of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)

²(Water Resources Research Institute, Xi'an University of Technology, Xi'an 710048, China)

³(College of Economics and Management, Xi'an University of Technology, Xi'an 710048, China)

Abstract: The sudden water pollution incidents occur frequently, so it has become a trend to do simulated research on sudden water pollution incident utilizing GIS platform, remote sensing technology and water quality model. Aiming at solving the problem that mass, heterogeneous and dynamic remote sensing data could not be disposed rapidly and stored efficiently; the storage system of remote sense tile data used in the sudden water pollution incidents was designed and implemented based on HBase. The system is supported by the digital earth platform and integrated with remote sensing tile pyramid model and MapReduce framework. The linear quad tree coding index of the remote sensing tile data was improved in this system according to the frequent load character of remote sensing tile data caused by event situation with real-time and dynamic observation and stored into HBase together with water pollution data. The access efficiency of the remote sensing tile data was increased through adding cache area. Applying this system to Weihe River sudden water pollution in Shanxi, the results showed that the improved index could extract the remote sensing data of the water pollution area within 22 milliseconds and the introduction of the caching mechanism saved almost 2/3 loading time on the remote sensing data in water pollution area. Therefore, the storage system could be accessed efficiently and it could satisfy the quick emergent simulation demand of the sudden water pollution incidents.

Key words: sudden water pollution incidents; remote sensing tile data; big data; MapReduce; HBase

① 基金项目:“十二五”国家水体污染控制与治理重大专项课题(2012ZX07201-006);陕西省自然科学基金研究计划(2015JQ5157);陕西省教育厅科学研究计划(14JK1293);西安工程大学博士科研启动基金(BS1330)

收稿时间:2015-05-06;收到修改稿时间:2015-06-08

突发水污染事件频发,给现代人类社会环境、经济发展成果造成了巨大冲击.为有效应对突发水污染事件,相关学者多采用数值模拟技术和可视化模拟技术对其研究^[1-3].传统数值模拟技术为突发水污染事件的处置提供了理论基础,可视化模拟技术可帮助决策者动态、直观地掌握污染物在水体中的运移过程,增强决策者同水污染事件的交互性,进而防止污染事态的进一步扩散.尤其是利用GIS平台、遥感技术结合水质模型对水污染事件进行研究已逐渐成为一种发展趋势.然而,随着卫星遥感技术的发展,高分辨率影像数据正在成几何级数增长,更多地表现出海量、多源、异构、动态等大数据特征,如何结合突发水污染事件的应急应用特点,对海量遥感影像大数据进行快速处理、存储便成为能否有效应对突发水污染事件的一大瓶颈.

近几年来,大数据、分布式、云计算等现代信息技术为大规模、高精度、多尺度、异构、动态、海量遥感数据的处理、存储和应用提供了新的途径.如文献[4]以文件目录方式对碎片数据进行分布式存储,解决了整体遥感影像数据的共享问题.霍树民^[5]基于MapReduce对批量遥感影像瓦片金字塔进行了并行构建,提出了基于MapReduce和HBase的海量影像数据并行入库的算法,缩短了海量影像数据入库的时间.文献[6]设计了云计算环境下的高分辨率遥感影像存储模型、管理平台及高性能计算等服务,其中采用了Hadoop中HDFS和HBase技术对高分辨率遥感影像数据进行了存储.夏英^[7]教授则结合金字塔模型和MapReduce架构提出了一种适用于云计算环境下的分布式并行存储BML(storage based on MapReduce and local file system)方法,该方法基于金字塔模型对影像数据进行分层分块处理,并对所得瓦片重新编码,其与传统数据存储方法相比,在海量影像数据存储性能方面有一定提高.为了更好地满足海量数据库高并发读写、高效存储和访问、高扩展性和高可用性的需求,文献[8]基于MongoDB的NoSQL类数据库对地图瓦片数据进行了存储,此方案与采用关系型数据库相比,具有明显的高效性.

由于大数据技术等给行业的应用和发展带来了便利,因此,水利行业的多位学者应用该技术进行了较为深入的研究.如文献[9-12]就分别采用大数据技术对多类型水利数据的存储、组织、交换、共享、挖掘

和应用进行了分析,提出或建立了支撑水利大数据的关键技术、数据中心、处理模式等,进而提高水利事务与业务的决策支持能力,为流域突发水污染事件提供了预测,也为推进水利大数据资源化进程进行了有益地探索.尤其是孟令奎^[13]教授等采用大数据技术中的HDFS文件系统和MapReduce并行计算框架实现了空间数据密集计算、并行计算、快速检索与快速传输等关键技术,解决了海量多源遥感数据的低处理效率和共享程度低等问题,为水利遥感数据中心的建设提供了参考和建议.

尽管大数据技术给人们处理海量遥感数据提供了技术支撑,然而,突发水污染事件发生后,决策者更加关注事件发生地影像数据的加载速度,从而为决策者提供更加快捷的决策服务,但目前关于影像数据的处理研究中还存在以下问题:①考虑了对遥感影像数据的并行入库、定位检索、移动操作等问题,但大部分文献将这些问题割裂开来,特别在设计影像数据入库过程中没有考虑到其同定位、缩放、平移等操作之间的关系;②在遥感数据的处理过程中,相邻数据的处置可提高影像加载的效率,但考虑相邻数据时,仅考虑了同一层间相邻数据之间的关系,而忽略了研究区域遥感影像数据放大、缩小等操作对高效加载数据的需求.因此,本文在分析遥感大数据的基础上,结合突发水污染事件的特点,建立一个可快速存储和访问相邻影像数据,并在不同遥感影像图层之间进行切换的系统,最后以渭河陕西段水污染事件为例进行分析,结果表明,本文建立的遥感数据存储系统设计良好,能够快速、有效地处理海量遥感数据,从而为突发水污染事件的应急决策提供了技术支撑.

1 水污染遥感瓦片大数据存储系统

突发水污染数据覆盖范围已从简单的水环境、水文方面扩大至水利、气象、工程、管理、决策、生活等各个方面,然而作为其中应对突发水污染事件的一项重要重要的遥感数据,其数据量在不断的变化,具有多源、异构、海量、动态等特点.本文针对水污染事件的这些特点,设计和实现了基于HBase的突发水污染遥感瓦片大数据存储系统,该系统以数字地球平台为支撑,结合遥感瓦片金字塔模型、MapReduce编程模型,按照水污染事件对遥感瓦片数据的操作类别,改进并实现了瓦片数据的存储索引技术,将其与水污染

数据一同存入到 HBase 分布式数据库中, 并通过增加缓存区提升瓦片数据的访问速率, 为决策者定位、平移等操作及快速应对水污染事件提供了技术支持. 该存储系统如图 1 所示.

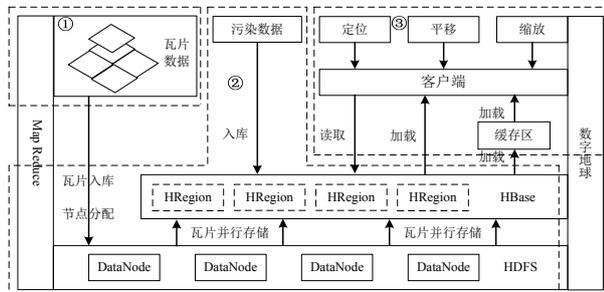


图 1 基于 HBase 的水污染遥感瓦片大数据存储系统

该系统包含多分辨率遥感瓦片金字塔模型的构建、遥感瓦片数据的并行入库及遥感瓦片数据的快速加载 3 部分内容. 遥感瓦片金字塔模型的构建是应对突发水污染事件的基础, 为遥感数据的瓦片化存储和处置提供了便利; 遥感瓦片数据的并行入库是借助 MapReduce 框架, 利用 Hadoop 架构将瓦片数据依次快速分配至 HDFS 的各个节点上, 同时并行化存入到 HBase 数据库中, 改变传统的串行入库模式; 遥感瓦片数据的快速加载依靠的是良好的瓦片索引的设计, 该索引不仅考虑了同层瓦片数据之间的相邻关系, 还考虑了瓦片数据的层间相邻关系, 并通过引入缓存机制来提升瓦片数据加载效率, 从而增强事件应急响应的可视化模拟或决策效果.

2 关键技术

2.1 遥感瓦片金字塔模型及其四叉树编码

水污染事件发生后, 需要运用可视化等技术对污染团的运移过程进行展现, 有针对性地快速调度受污染地区的遥感数据, 从而达到高效处置水污染事件的目的. 然而, 由于目前遥感影像数据采集手段和方法的多样性, 导致了遥感数据具有格式不统一、存储空间大、分布式存储等特点, 所以本文采用瓦片金字塔技术对遥感数据进行管理和组织.

瓦片金字塔技术是一种多分辨率层次模型, 通常在构建瓦片金字塔时总是采用倍率法构建, 形成分辨率从低到高的多个分辨率层次. 在瓦片金字塔模型下, 同一范围的数据是分层显示的, 同层数据又被分割成大小相同的小块, 可以根据实际需要人为的规定层与

层之间的分辨率变化倍率, 从而确定所分层数和每层所分块数. 瓦片金字塔模型中最小的单元被称为瓦片影像或瓦片, 它们是每层分割出来的数据块, 是空间影像数据的基本单位. 由于直接存储遥感瓦片数据会耗费大量的时间, 为此利用线性四叉树便于检索的优点对遥感瓦片进行编码. 瓦片金字塔模型及其与线性四叉树编码之间的对应关系如图 2 所示.

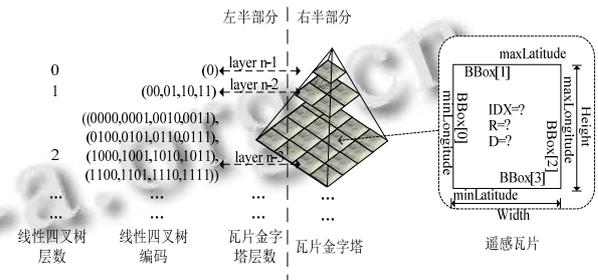


图 2 瓦片金字塔模型及其与线性四叉树编码的对应关系

图 2 左半部分是线性四叉树编码, 图 2 右半部分为瓦片金字塔模型. 图 2 左半部分中是一个具有 n 层图像的影像金字塔结构, 其第 i 层四叉树编码层和影像金字塔的第 $n-i$ 层对应, 而影像金字塔的构建便是对从第 0 层(瓦片金字塔的最低层)高分辨率的图像进行采样抽取, 形成第 1 层次高分辨率的瓦片, 依次循环直到第 $n-1$ 层低分辨率瓦片图像的生成. 所以从影像金字塔的第 1 层开始直到第 $n-2$ 层, 第 i 层图元均是由下一层的四个相邻的四叉树编号前 $2*(n-i-1)$ 位数相同的瓦片组成. 图 2 右半部分中金字塔各层分辨率的公式为: $R_n = R_0 \times \lambda^{-n}$, 其中 R_0 为数据原始分辨率, λ 为倍率, $d(d=0,1,\dots,17)$ 为数据所属层数, 一共有 $n+1(n=17)$ 层. 其中, 瓦片由一个六元组 $\{IDX, D, R, W, H, BBOX\}$ 表示. IDX (Index) 是一个唯一的编码符号, 用于描述该瓦片的空间索引; D (Depth) 表示瓦片所在的塔层位置; R (Resolution) 为当前瓦片的分辨率; W (Width) 和 H (Height) 则用以表示瓦片的宽度和高度(幅度大小); $BBOX$ (Bounding Box) 则给出影像的空间位置, 它由 $\{minLon, minLat, maxLon, maxLat\}$ 四个元素表示, 简记为 $BBOX = \{iLo, iLa, xLo, xLa\}$, 分别对应于空间区域的最小经度、最小纬度、最大经度和最大纬度, 并记 $spanLon = xLo - iLo$, $spanLat = xLa - iLa$ 分别用来表示瓦片的经度跨度和纬度跨度^[14].

2.2 改进型索引遥感瓦片数据存储模型

为了能够快速地查找、定位并加载遥感瓦片数据, 本文采用文献[15]中影像金字塔和线性四叉树之间的映射关系, 设计了基于 HBase^[16]数据库的以 Key-value 格式存储的瓦片影像存储模型, 如图 3 所示.

Row Key	Time Stamp	Column Family								
		Image Block	D	R	W	H	BBOX	T	C	
r ₁	t ₃	block _{1,β}	1					T ₃	25	HRegion
	t ₂	block _{1,ε}	1					T ₂	10	
	t ₁	block _{1,γ}	1					T ₁		
r ₂	t ₂	block _{2,ε}	2							HRegion
	t ₁	block _{2,γ}	2							

图 3 基于 HBase 的遥感瓦片数据存储模型

图 3 中, RowKey 值为 $r_i (i=1,2,\dots)$, 为影像瓦片的索引信息, 和图 2 中的 IDX 对应; TimeStamp 为时间戳, 用 $t_j (j=1,2,\dots)$ 表示, 是遥感数据进入到 HBase 数据库中的时间; Column Family 列用来保存瓦片影像数据及其对应的属性信息, Image Block 列保存实际的瓦片影像数据; D 为树的层高或深度; R 为当前瓦片影像的分辨率; W 和 H 分别为瓦片影像的宽度和高度; $BBOX$ 表示瓦片数据的经纬度范围, 由 $\{iLo, iLa, xLo, xLa\}$ 组成, 分别对应瓦片区域的最小经度、最小纬度、最大经度、最大纬度; T 为遥感数据实际拍摄的时间; C 为污染物浓度值, 单位为 mg/L .

采用四叉树方式编码, 除顶层遥感数据的编码为 0 外, 第 $d (d=1,2,\dots,17)$ 层的编码为 $EC=H_1H_2\dots H_{d-1}H_d$. 其中, $H_d \in \{S_dW_d, S_dE_d, N_dW_d, N_dE_d\}$, S_dW_d 、 S_dE_d 、 N_dW_d 、 N_dE_d 分别表示第 d 层四叉树遥感瓦片中的西南、东南、西北、东北方向上的瓦片; S_d 和 W_d 编码为 0, N_d 和 E_d 编码为 1, S_dW_d 、 S_dE_d 、 N_dW_d 、 N_dE_d 的编码分别为 00、01、10、11. 因此, 第 d 层的编码可表示为 $m_1m_2\dots m_{2^{d-1}-1}m_{2^d}$, $m_j \in \{0,1\}$, $j=1,2,\dots,2^d-1,2^d$. 此时, 每个遥感瓦片均可用编码 EC 来表示, 和 r_i 相对应, 作为最初瓦片数据存储模型中的索引值, 这为瓦片数据的存储和普通查询操作提供了便利.

然而, 突发水污染事件发生后, 决策者在关注事态进展时, 短时间内会对发生区域的遥感瓦片数据有大量的需求, 不仅仅有瓦片数据的定位和普通查询, 更多的是对数字地球进行平移、缩放等操作引起的邻

近查询等需求, 而现有瓦片索引技术在设计过程中并未考虑这些因素, 因而对突发水污染事件的应急效率影响较大. 为此, 本文将遥感瓦片 HBase 数据库中的多列进行组合作为查询条件, 将其作为“key-value”格式的 key 值, 通过 key 值引擎快速查找目标瓦片数据. 所以, 改进后的 RowKey 值为 $rt_EC_d_ilo-xlo-ila-xla_T$, 其中 rt 为索引类别, 其他符号含义同上, 该索引值充分考虑了索引类别、瓦片四叉树编码、瓦片层级、瓦片经纬度范围和遥感数据拍摄时间等信息. 根据水污染事件的特点, 本文的索引类别只包含 2 种情况: 按照瓦片编码查询和按照层级同经纬度联合查询, 在配置文件中用 type 表示^[17] (如图 4 所示). 其中, type 为 tt 时表示需要平移、缩放等操作; type 为 pt 时表示需要按照经纬度和层级进行定位等操作.

```

12 <index>
13 <type>tt</type>
14 <table>tilePyramid</table>
15 <fields>
16 <field qualifier="EC"/>
17 <field qualifier="T"/>
18 </fields>
19 </index>
20 <index>
21 <type>pt</type>
22 <table>tilePyramid</table>
23 <fields>
24 <field qualifier="d"/>
25 <field qualifier="ilo"/>
26 <field qualifier="xlo"/>
27 <field qualifier="ila"/>
28 <field qualifier="xla"/>
29 <field qualifier="T"/>
30 </fields>
31 </index>
    
```

图 4 部分配置文件

水污染事件发生后, 根据对瓦片数据的操作定制查询请求, 分析查询类别后和 RowKey 索引值中的 rt 进行比对, 找到合适的索引, 根据索引加载需要的遥感瓦片数据, 本文定义的 2 种操作如下: ①当初次定位水污染发生地时, 输入发生地名后可按照地区经纬度信息(lon, lat)自动定位到该地区(默认层级是 16 层), 此时查询条件可表示为 $Condition.and(Condition.and(Condition.condition("d", Operator.EQ, "16"), Condition.condition("T", Operator.EQ, t)), Condition.and(Condition.and(Condition.condition("ilo", Operator.LT, lon), Condition.condition("xlo", Operator.GT, lon)), Condition.and(Condition.condition("ila", Operator.LT, lat), Condition.condition("xla", Operator.GT, lat))))$, 根据查询结果可得到遥感瓦片及其对应的编码 EC 值; ②当定位到该地区后, 利用 EC 值编码方法相较于其他 BSP 树、K-D-B 树、R 树等在查

找邻域瓦片数据应用上具有较强的优势,在数字地球平台上进行平移、缩放等操作,从而快速地响应水污染应急决策事件。

2.3 海量遥感瓦片数据的快速存取流程

海量遥感瓦片数据的快速存取流程描述如图 5 所示,包含遥感瓦片数据的并行入库和引入缓存机制的瓦片加载 2 部分内容 3 个步骤:①瓦片数据入库;②瓦片数据查询;③瓦片数据加载。

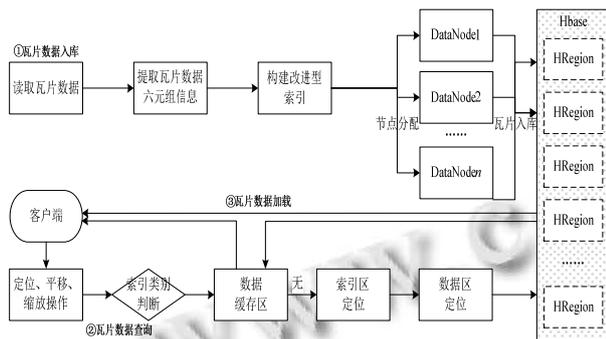


图 5 海量遥感瓦片数据的快速存取流程

2.3.1 遥感瓦片数据的并行入库

传统的瓦片数据是直接存入到数据库中,或者存储到分布式环境中的各个节点上,然而随着全球海量瓦片数据的出现,此种方式已经不能适应水污染应急事件的应用需求,因此,本文采用 MapReduce 框架将瓦片数据以 Block 大小分层分配到 DataNode 节点上.在存放过程中,应用 Hash 算法对瓦片索引 RowKey 进行计算,使瓦片数据尽可能均匀地分布在各个节点上,从而防止瓦片数据在读取或入库过程中因某台服务器负载过高而造成整体性能瓶颈.数据分配到 DataNode 节点后,再通过 Map 过程将本地分布式文件系统的数据导入到包含分布式数据库 HBase 的 HRegion 上.其中 Hash 算法如下:根据 HBase 遥感瓦片数据存储的索引规则,取影像瓦片的二进制编码索引信息 r_i ,将其转换为十进制数,然后对其取 Hash 编码,当 RowKey 的哈希值正好等于数据服务器编号,则将该遥感瓦片数据直接存放在该数据库服务器上;若该数据库服务器已存满数据,则寻找第一个大于该数据库服务器编号的 DataNode 节点,直到所有瓦片数据存完。

在瓦片数据存储过程中,还应根据遥感数据总存储量、瓦片数据大小、Block 大小来确定 DataNode 节点个数.如对于空间分辨率为 $X \times Y$ 的遥感影像来讲,

若构建一个 n 层金字塔模型,每层分辨率缩小 K 倍,剖分后的空间分辨率为 $R \times R$,总瓦片数 $num_t = \sum_{i=1}^n ((X \times Y) / (R^2 \times K^2 \times i))$,则存储这些瓦片需要的 Block 数目为 $num_b = \sum_{i=1}^n [X / (R \times K \times i)] \times [Y / (R \times K \times i)] \times S / L$.其中, S 为剖分后瓦片影像数据大小, L 为 Block 大小.假设每个 DataNode 的大小相同且至多包含 DB 个 Block,则这些瓦片数据至少需要 $\lceil num_b / DB \rceil$ 个 DataNode.

2.3.2 引入缓存机制的瓦片加载

瓦片加载包含瓦片数据查询和瓦片数据加载 2 个过程.为快速加载遥感瓦片数据,本文引入缓存机制.当数字地球客户端首次发出地图请求服务时,将显示区域瓦片和其相邻的瓦片数据一同载入到缓存中,并在客户端加载显示区域数据.当用户进行放大、平移等操作时,可先去缓存区取数据,若缓存区没有匹配的数据,再到 HBase 数据库中搜寻和待加载区域相邻的瓦片数据.缓存区加载瓦片数据的多少依赖于两个关键因素:缓存区容量的大小和通过数字地球客户端对水污染区操作的频繁程度。

3 实例应用及结果分析

本文首先采用 NASA 的 BMNG 影像数据构建了一个全球多分辨率虚拟地形环境,生成不同级别分辨率的金字塔影像 320 000 多幅,总计 6.71GB.然后用 5 个 IBM System x3650 机架式服务器搭建了集群环境,硬件参数为:处理器 Intel Xeon E5-2609 2.4GHz,内存 4GB,硬盘 300G;其中 1 个为主节点,4 个为子节点,主节点运行 NameNode 和 JobTracker,子节点运行 DataNode 和 TaskTracker,IP 地址分别为 202.200.115.100~202.200.115.104.软件版本:操作系统版本为 rhel-server-6.4, JDK 版本为 Oracle JDK 1.6.0_05, Hadoop 版本为 1.2.1, HBase 版本为 0.94.1.考虑到遥感瓦片容量及服务器硬件参数的限制,实验中将 Map 的任务数设置为 20, Reduce 任务数设置为 5,即每个 TaskTracker 节点同时运行 4 个 Map 任务及 1 个 Reduce 任务.为有效应对突发水污染事件,遥感瓦片数据的处理分两步走.第一步是全球瓦片影像数据的瓦片化及入库,因为遥感瓦片数据入库过程非常耗时,所以第一步必须在污染事件发生前定期进行更新入库;第二步是为查看水污染区域污染物的运移态势,在数字地球平台上进行定位、平移、缩放等操作。

由于渭河是黄河第一大支流，在陕西段全长 502 千米，流域面积 6.72 万平方公里，占陕西全省总面积的三分之一，因此本文以陕西省境内渭河发生水污染事件为例。假设某大型石油公司因突发事件造成大量石油从渭河陕西段上游宝鸡县境内流入河流，造成下游水体污染，通过对数字地球平台的操作可快速模拟并预测到流域下游地区石油的运移和扩散状态(如图 6 所示)。为验证本文所建遥感瓦片大数据的存储系统，特设计 2 个实验环境。



图 6 渭河陕西段上游水污染事件发生区域

3.1 原索引与改进后索引对瓦片数据存取效率的影响

索引设计的好坏会影响遥感瓦片数据的入库和加载效率，为此，本文对采用线性四叉树编码方式的原索引与改进后的索引进行了对比，结果如图 7 所示(说明：由于瓦片入库时间在百秒级别，而瓦片加载时间在百个毫秒级别，为统一起见，图 7 中 2 种操作的时间单位不一致。)

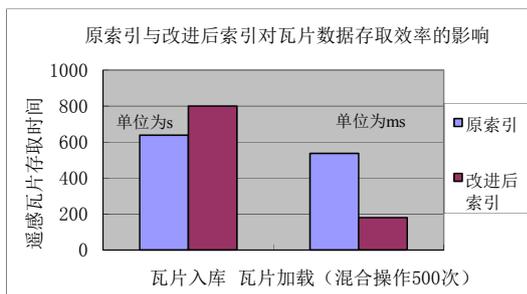


图 7 原索引与改进后索引对瓦片数据存取效率的影响

从图 7 中可以看出，在瓦片入库时，改进后索引比原索引耗费的时间更多，这是由于瓦片入库时索引改进过程会增加一些时间；尽管如此，良好设计的索引为瓦片加载节省了近 2/3 半的时间。因此改进型索引在提取突发水污染事件发生区域遥感数据时具有较

强的优势。

3.2 缓存大小对水污染区域遥感数据加载效率的影响

增加缓存区可提升遥感瓦片数据的加载效率，若缓存区太小，则相邻的瓦片数据不能全部加载到缓存区，而如果要将所有瓦片数据全部加载到缓存区，则读取出来的瓦片数据会占用很大的内存，这两种情况均会影响到遥感数据的加载效率。为此，本文以水污染区域为研究对象，通过缓存区大小的设置来研究平移、缩放及混合操作对遥感瓦片数据加载效率的影响，结果如图 8 所示。

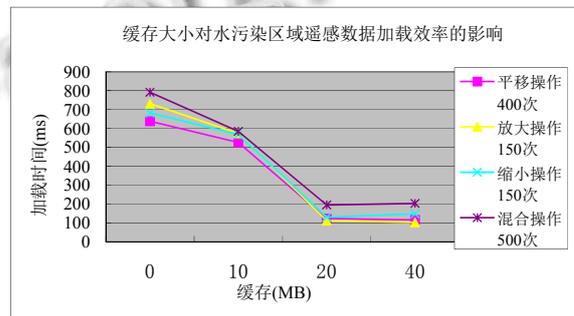


图 8 缓存大小对水污染区域遥感数据加载效率的影响

从图 8 中可以看出，当不设置缓存区时，各种操作对瓦片数据的加载的耗时均较高，随着缓存区容量的增大，遥感瓦片数据的加载效率在 20MB 时处于一个临界状态，即增大缓存区容量后，所有操作在加载遥感数据效率方面没有太大的改善。因而缓存区大小对水污染事件的可视化模拟会有一些影响，因而需要结合瓦片数据量大小进行合理设置。

4 结语

本文主要针对突发水污染事件的应用方向及遥感瓦片数据的特点，设计了基于 HBase 的水污染遥感瓦片大数据存储系统，构建了多分辨率遥感瓦片金字塔模型，对遥感瓦片数据存储的索引进行了改进，体现了在数字地球客户端进行定位、平移、缩放等操作对瓦片数据加载效率的影响，实现了遥感瓦片数据的并行入库和引入缓存机制的瓦片加载方案。通过实例证明，本文设计的遥感瓦片大数据存储系统具有高效的访问效率和扩展性，但后续还有一些工作需要完善。

1)在设计基于 HBase 的遥感瓦片数据库中尽管增加了水污染浓度值，但由于水污染事件复杂、模型复杂，因此需要将 Hadoop 框架、HBase 数据库同水污

染模型结合起来,在更大程度上发挥Hadoop框架的优势.

2)本文采用了Hash算法将海量遥感瓦片数据均匀地分布在各个节点上,尽管可以降低因数据频繁访问导致某一节点性能降低的影响,但这种算法弱化了数据之间的联系,而瓦片数据之间在某一种程度上是有联系的,因此在后续工作中还应着重研究瓦片数据之间的局部性,从而有效提升海量瓦片数据在局部范围内的访问性能.

参考文献

- 1 吴迪军,黄全义,孙海燕,等.突发性水污染扩散模型及其在GIS平台中的可视化.武汉大学学报(信息科学版),2009,34(2):131-134.
- 2 吴迪军,陈建国,黄全义,等.水污染扩散的二维数值模拟及其可视化.武汉大学学报(工学版),2009,42(3):296-300.
- 3 李文君,黄强,杨明祥.松花江突发水污染可视化模拟系统研究与实现.计算机工程与应用,2012,48(9):211-214.
- 4 陈田庆,解建仓,李建勋,等.分布式环境下的遥感影像组织与调度.计算机工程,2010,36(23):9-12.
- 5 霍树民.基于Hadoop的海量影像数据管理关键技术研究[硕士学位论文].长沙:国防科技大学,2010.
- 6 康俊锋.云计算环境下高分辨率遥感影像存储与高效管理技术研究[博士学位论文].杭州:浙江大学,2011.
- 7 夏英,杨选伦.云环境中基于金字塔模型的影像数据存储方法.重庆邮电大学学报(自然科学版),2012,24(6):669-674.
- 8 陈超,王亮,闫浩文,等.一种基于NoSQL的地图瓦片数据存储技术.测绘科学,2013,38(1):142-143,159.
- 9 莫荣强,艾萍,吴礼福,等.一种支持大数据的水利数据中心基础框架.水利信息化,2013,31(3):16-20.
- 10 冯吉平,陈微,官涤,等.大数据技术在松辽流域水环境管理中的应用展望.水利发展研究,2014,14(9):63-65,71.
- 11 陈龙,万定生,顾昕辰.基于Hive的水利普查数据仓库.计算机与现代化,2014,30(5):127-130.
- 12 冯钧,许潇,唐志贤,等.水利大数据及其资源化关键技术研究.水利信息化,2013,31(4):6-9.
- 13 李继园,孟令奎,田洋,等.水利遥感数据共享平台的高性能运行机制研究.水利信息化,2011,29(2):9-14.
- 14 李建勋,沈冰,郭建华,等.面向影像金字塔的线性二叉树编码及其特性.计算机应用,2011,31(4):1148-1151.
- 15 李建勋,沈冰,姜仁贵,等.面向影像金字塔的四叉树空间索引算法.计算机工程,2011,37(10):11-13.
- 16 Apache. <http://hbase.apache.org/>. [2015-1-1].
- 17 耿立超.HBase 高性能复杂条件查询引擎.<http://blog.csdn.net/bluishglc/article/details/31799255>. [2014-6-17].