计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
49(Suppl.): 12-18, 2012

基于 Hadoop 的海量数据存储平台设计与开发

崔 杰[1] 李陶深[1] 兰红星[2]
[1](广西大学计算机与电子信息学院 南宁 530004)
[2](广西工业和信息化委员会 南宁 530022)
(cuijietianlong@163.com)

Design and Development of the Mass Data Storage Platform Based on Hadoop

Cui Jie[1], Li Taoshen[1], and Lan Hongxing[2]
[1](*School of Computer, Electronics and Information, Guangxi University, Nanning* 530004)
[2](*Guangxi Industry and Information Technology Committee, Nanning* 530022)

Abstract With the development and utilization of BeiBu Bay Marine ecological resources, mass marine science data rapidly emerge in large numbers and it is very important to use a mass data storage platform to manage and store these science data reasonable. This paper puts forward the management and storage the mass marine science data methods based on the distributed computing technology, builds the mass marine science data storage platform solutions, designs and develops a mass data storage platform based on Hadoop by using Linux cluster technology. This system which consists of five modules includes system management module, parallel loading storage module, parallel query module, data dictionary module, backup and recovery module and it can achieve to store massive amounts of marine science data. The system module achieving result shows that this system enjoys good safety, reliability, easy maintenance and good expansibility.

Key words mass data storage; marine science data; Hadoop; distributed computing

摘 要 随着北部湾海洋生态资源的开发和利用,海量海洋科学数据飞速涌现出来,利用海量数据存储平台合理管理和存储这些科学数据显得极为重要.这里提出了一种基于分布式计算技术进行管理和存储海量海洋科学数据方法,构建了海量海洋科学数据存储平台解决方案,采用 Linux 集群技术,设计开发一个基于 Hadoop 的海量数据存储平台.系统由五大模块组成,有系统管理模块、并行加载存储模块、并行查询模块、数据字典模块、备份恢复模块,能够实现存储海量海洋科学数据.系统模块实现结果表明,该系统安全可靠、易维护、具有良好的可扩展性.

关键词 海量数据存储;海洋科学数据;Hadoop;分布式计算

中图法分类号 TP311.13

随着《北部湾经济区发展规划》颁布实施,以北部湾经济区海洋为研究样本的系列重大基础研究专项和重大科学研究项目正在逐一展开,届时将产生海量的海洋科学数据,这些数据具有海量、复杂、多样、异构、动态变化等特性.而且目前各项目的海洋科学数据均缺乏统一的采集和存储的标准及规范,

收稿日期:2012-01-04
基金项目:国家自然科学基金项目(60963022);广西自然科学基金重点项目(桂科自 0832056);广西大学拔尖创新团队建设计划项目(L300249);广西研究生教育创新计划项目(GXU11T32550)

形成“数据孤岛”. 如何存储和管理海量的海洋科学数据,使这些数据得到高效的利用,成为进行海洋科学研究项目的关键之一. 因此构建一个北部湾海洋科学数据存储平台是目前充分发挥各重大基础科学研究项目研究效益的现实途径,也是北部湾经济区可持续发展的必然要求.

传统的对大规模数据处理大多使用分布式的高性能计算、网格计算等技术,需要耗费昂贵的计算资源,而且对于如何把大规模数据有效分割和计算任务的合理分配都需要繁琐的编程才能实现,而 Hadoop 分布式技术的发展正好可以解决以上的问题. Hadoop 是 Apache 开源组织的一个分布式计算框架,可以在大量廉价的硬件设备组成的集群上运行应用程序,构建一个具有高可靠性和良好扩展性的并行分布式系统,Hadoop 分布式文件系统(Hadoop Distributed File System, HDFS)、MapReduc 编程模型和 HBase 分布式数据库是其三大核心技术[1-3]. 本文在 Linux 集群技术的基础上,利用 Hadoop 分布式技术,对北部湾海量海洋科学数据高效的处理后存储到可扩展的分布式数据库中,设计并实现一个易扩展的高效的海量数据存储管理系统.

1　平台总体设计

1.1　平台总体框架结构

结合海量数据异构性、分布性、多样性等特点,从系统编程实现角度考虑,本系统采用 MVC 3 层架构设计,使结构更加清晰,系统易于扩展. 系统整体架构如图 1 所示:

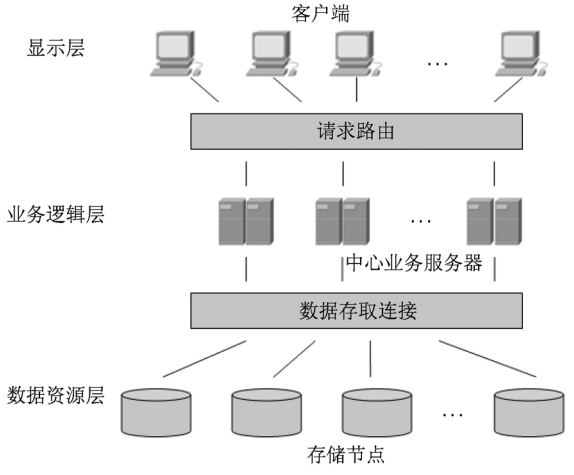


图 1　平台整体框架结构

显示层:为用户提供方便、易用和友好界面,普通用户可以通过页面浏览和查询海洋数据,高级用户可以利用系统提供的公共 API 接口,扩展系统[4].

业务逻辑层:并行处理海量海洋科学数据,并对整个平台系统配置管理.

数据资源层:是整个平台的基础,存储和管理海量海洋科学数据.

1.2　平台总体功能设计

从系统功能角度考虑,可以将整个系统分 3 层,如图 2 所示:

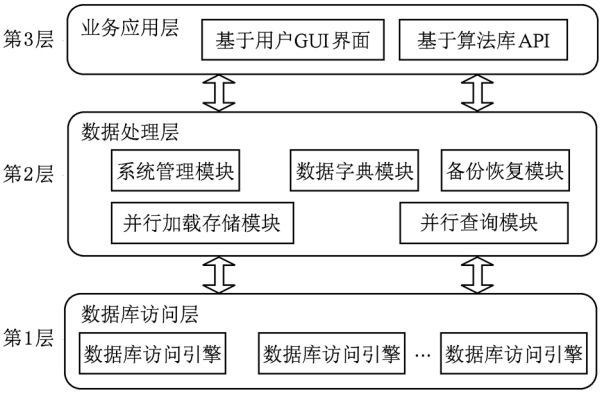


图 2　系统功能分层设计

第 1 层是数据访问层. 对于海量数据存储,在存取数据时不会只局限对一种数据库的操作,本层需要对各种数据库提供的不同数据源进行屏蔽,提供数据库访问服务,这样系统才能够适应处理存储海量数据的要求,具有较好的可扩展性和完备性,方便管理和部署.

第 2 层是数据处理层. 数据处理层作为整个系统的核心,同时也是本系统设计开发的重点内容. 它采用分布式数据库技术、Linux 集群技术等,提供了对海量数据的并行加载存储等主要功能[5]. 该层通过对海量数据并行处理,把处理后的数据存储到本系统的分布式数据库中,同时还提供了保证系统能够正常运行的管理支撑服务.

该层分为 5 个功能模块:系统管理模块、并行加载存储模块、并行查询模块、数据字典模块、备份恢复模块.

1) 系统管理模块又分为:负载平衡管理、系统日志管理、对象事务管理、系统远程部署管理、自主维护管理等. 系统管理模块:对系统实现分布式管理. 负载平衡用于存储节点的负载均衡和容错管理;日志管理用户记录系统运行的运行轨迹、关键事件

和状态记录等;对象事务管理用来对系统事务处理及其一致性进行管理;系统远程部署管理用来对远程集群的部署和配置,实现系统整体最优运行状态;自主维护管理用来对系统自身运行状态的监测,根据具体运行状态进行自我调整. 以上功能在 Hadoop 基础平台中可以通过合理配置其组件来实现,让各个组件协同工作达到最优.

2) 并行加载存储模块又分为:并行数据加载模块、并行 ETL 模块、并行存储模块等. 如图 3 所示:

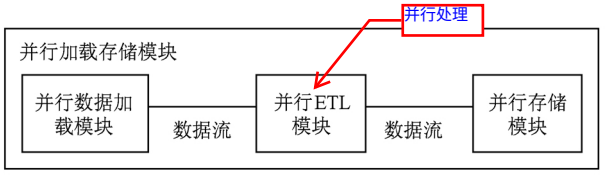


图 3 并行加载存储模块组成

并行加载存储模块:提供对海量数据的并行加载、处理和存储功能. 并行加载模块将数据从其他外设中导入平台的 HDFS;并行 ETL 模块用来对 HDFS 中的原始数据进行处理得到存储数据;并行存储模块提供对处理后的数据进行存储[6].

3) 并行查询模块:提供对海量数据的并行查询、用户自定义事务处理等功能.

4) 数据字典模块:为系统配置一个全局的数据字典,用于维护并行数据库的元数据信息.

5) 备份恢复模块:提供对系统存储数据的备份管理、备份存储、备份恢复等功能,增强系统的安全性和容错性.

第 3 层是业务应用层. 分为基于用户 GUI 界面和基于算法库 API.

1) 基于用户 GUI 界面:用户可以通过简单应用的操作界面工具,进行海量数据处理存储.

2) 基于算法库 API:对于高级用户可以编写应用系统,调用算法库中的 API 来扩展本系统,实现所需的应用功能.

1.3 平台网络拓扑结构

从图 4 中可以看出平台由多个数据库服务器和应用服务器组成,这些数据库服务器可以在同一地域,也可分布在不同地域. 随着数据量的增大和应用需求的复杂变化,平台可以很容易的扩展,而这些变动对用户来说都是透明的,并且现有的关系型数据库系统也可以整合到该平台中,通过去异构化处理共同为用户提供存储服务[7],从而为用户透明地提供存储和管理海量海洋科学数据的功能. 该平台可以安全、稳定、不间断的为政府、企业、个人等用户提供海量数据存储服务,使北部湾海洋科学数据能够得到妥善的存储管理,更大地发挥其研究利用价值,为北部湾经济建设服务.

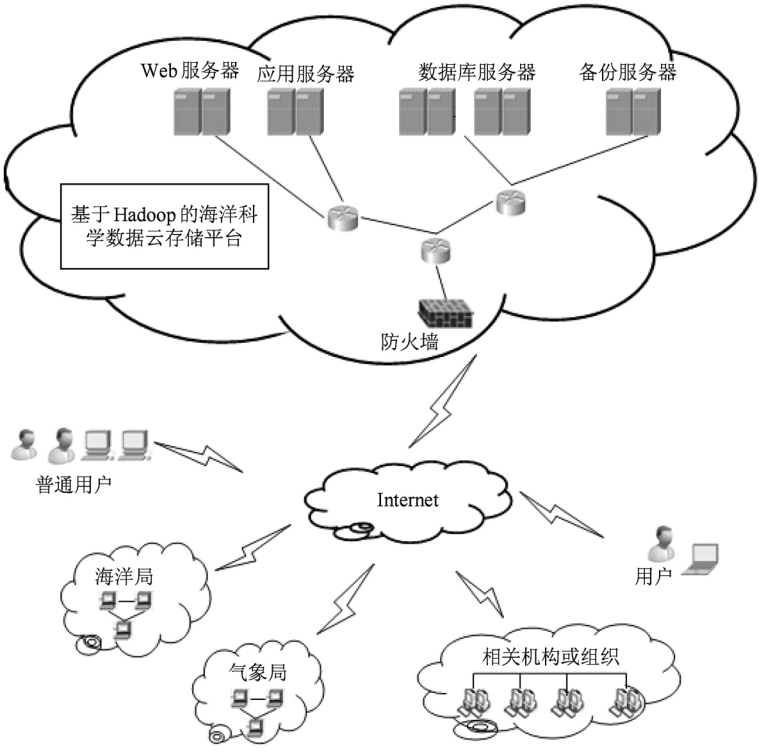


图 4 平台网络拓扑结构图

1.4　平台数据库整体设计

结合北部湾海洋自身特点,分析其海量数据的结构类型,对北部湾海洋科学数据存储平台数据库建设做以下整体设计,主要包括 12 个类型数据库[8-13].

1) 北部湾沿海生态数据库. 包括:北部湾海洋植物物种数据库;北部湾大型底栖生物物种数据库;北部湾动物物种及标本数据库;名贵、珍稀、濒危海洋动物数据库;海洋经济鱼、虾、贝、藻数据库;海水养殖动物(或鱼类,包括:外来养殖物种)种质资源元数据库;海洋生物模式标本数据库等.

2) 北部湾沿海遥感地理空间数据库. 包括:资源卫星数据库;遥感卫星图像检索数据库;北部湾沿海地貌数据库等.

3) 北部湾沿海重点海域海底基础环境空间数据库. 全面及时记录北部湾沿海重点海域海底基础环境的基本情况,包括数字地形地貌、海底底质性质与分布特征,及海底沉积物粒度分析、矿物分析、地球化学分析、微古分析结果及海洋钻探计划(ODP)勘测信息等.

4) 海洋基础地理地图数据库. 内容包括:海岸线及滩涂,岛屿、礁石、浅滩,海域水深及地形(等深线),沿岸陆域水系,陆地地形(等高线),重要的居民地,交通网,境界线(国界、省市界、县界、领海基线等),地名和地理名称及相关要素注记,区域界线等.

5) 北部湾沿海海洋水文数据库. 包括:营养盐数据库;温盐深声学数据库;流速数据库;海面气象数据库;潮汐数据库;波浪数据库等. 提供水文循环、大气对流、闪电、恶劣天气等大尺度的数据共享,包括实时和历史的数据.

6) 北部湾沿海环境数据库. 包括:潮汐预报信息数据库;海洋台站数据库;海流资料数据库;海洋气象观测资料数据库;海洋环境质量数据库等.

7) 北部湾海洋经济数据库. 包括:北部湾海洋综合经济数据库;海洋水产数据库;海洋石油天然气数据库;海滨砂矿数据库;海洋盐业数据库;沿海造船数据库;海洋交通运输数据库;北部湾沿海旅游数据库等.

8) 北部湾海洋资源数据库. 包括:北部湾沿海海岛概况数据库;沿海主要港口码头泊位及吞吐能力数据库;沿海盐场资源数据库;海洋石油天然气资源数据库;海洋旅游资源数据库;海洋自然保护区数据库;潮汐能资源数据库;波浪能资源数据库;潮流能资源数据库和海底电缆管道资源数据库等.

9) 海洋标准数据库. 主要收集与海洋、水产相关的国家标准和行业标准,全部为 PDF 格式.

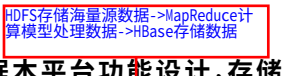
10) 海洋法规数据库. 收集与海洋、水产有关的中国法律、行政法规、部委规章、司法解释、地方法规、港澳台相关法律、国际条约及中共中央政策、其他机构文件库、判例案例、论文、合同范本等.

11) 中外文海洋数据库. 分水产养殖、水产品加工贮藏与综合利用、水生生物学、海洋生物学、海洋生物工程、海洋渔业、海洋化学、海洋环境与污染治理、海洋地质、海洋管理、物理海洋学等专题. 同时,也收录了中国大陆公开发表的海洋、水产方面的期刊论文,以全文的形式反映我国海洋、水产专业的学术发展水平.

12) 海洋音视频数据库. 收录与海洋、水产有关的音视频方面的文献,这些音视频文献可以使用通用的媒体播放器播放,并可提供下载服务.

目前数据收集工作正在进行中,数据库的建设也在同步进行,把海量科学数据经过处理存储起来统一管理.

2　海量数据存储平台开发



根据本平台功能设计,存储平台最主要的部分是数据处理层,而在实现数据处理层时,数据的并行加载存储模块成为了整个平台实现的核心,Hadoop 分布式技术为该平台提供了数据存储和数据处理的模型及方法[14-15]. 使用 Hadoop 分布式文件系统存储海量源数据,通过 MapReduce 分布式计算模型来处理这些海量源数据,然后采用 HBase 分布式数据库存储处理后的海量数据,以此来实现对海量海洋科学数据的存储管理.

2.1　Hadoop 分布式文件系统

HDFS 是分布式计算的存储基础,它具有高容错性,可以部署在廉价的硬件设备上,用来存储海量数据集,并且提供了对数据读写的高吞吐率[7-8]. HDFS 为北部湾海洋科学数据提供了海量存储的基础,作为未处理的源数据集保存在 Hadoop 分布式文件系统中.

HDFS 采用 Master/Slave 的体系结构,集群中由一个 NameNode 和很多个 DataNode 组成. NameNode 是主控服务器,管理文件系统元数据. 它执行文件系统的命名空间操作,比如打开、关闭、重命名文件或

目录,还决定数据块到 DataNode 的映射. DataNode 存储实际的数据,负责处理客户的读写请求,依照 NameNode 的命令,执行数据块的创建、复制、删除等工作. 一个集群只有一个 NameNode 的设计大大简化了系统架构. 体系结构如图 5 所示:

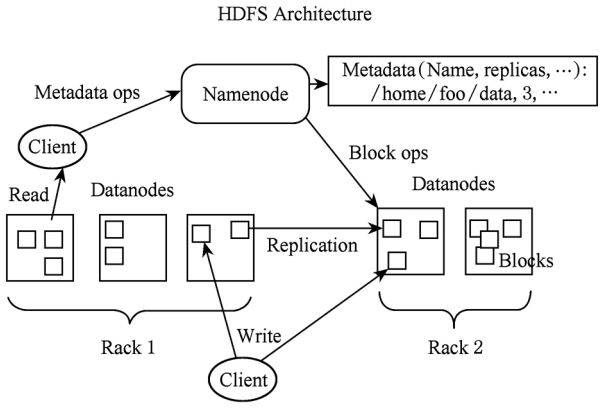


图 5 HDFS 体系结构

NameNode 使用事务日志(EditLog)来记录 HDFS 元数据的每次变化,使用映像文件(FsImage)存储文件系统的命名空间,包括数据块到文件的映射、文件的属性等等. 事务日志和映像文件是 HDFS 的核心数据结构. NameNode 启动时,它将从磁盘中读取映像文件和事务日志,把事务日志的事务都应用到内存中的映像文件上,然后将新的元数据刷新到本地磁盘新的映像文件中.

HDFS 还设计有特殊的 Secondary NameNode 节点,辅助 NameNode 处理映像文件和事务日志. 它会定期从 NameNode 上复制映像文件和事务日志到临时目录,合并生成新的映像文件后再重新上传到 NameNode,NameNode 更新映像文件并清理事务日志,使得事务日志的大小始终控制在某个特定的限度下.

2.2 MapReduce 编程

MapReduce 就是"任务的分解与结果的汇总". Map 把任务分解成为多个任务,Reduce 把分解后多任务处理的结果汇总起来,得到最终结果. 把从 HDFS 中读取的待处理的海量海洋科学数据分解成许多小数据集,每一个小数据集都并行处理,处理后存储到分布式数据库. 归纳如下:数据集分割$\langle k_1, v_1\rangle$map$\langle k_2, v_2\rangle$combine$\langle k_2, \text{list}(v_2)\rangle$reduce$\langle k_3, v_3\rangle$结果输出. 计算模型如图 6 所示.

将海量海洋科学数据分割 M 个片段进行并行 Map 操作,然后形成中间态键值对$\langle k, value\rangle$,接着

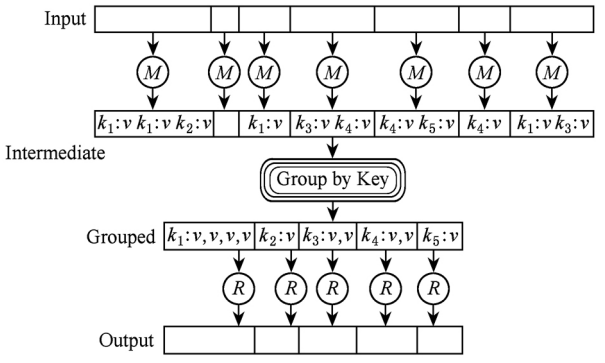


图 6 MapReduce 计算模型

以 k 值进行 Group 操作,形成新的$\langle k, \text{list}(value)\rangle$元组,对这些元组分割成 R 个片段进行并行的 Reduce 操作,最后输出到分布式数据库中保存起来. MapReduce 计算模型的实现是由 JobTracker 和 TaskTracker 这两类服务调度的. JobTracker 是主控服务器,只有一个,负责调度和管理 TaskTracker,把 Map 任务和 Reduce 任务分配给空闲的 TaskTracker,并负责监控任务的运行情况;TaskTracker 是从服务,可以有多个,负责执行任务. 通常 MapReduce 和 HDFS 是运行在一组相同的节点上的,即计算机点和存储节点通常在一起,方便高效地调度任务.

2.3 HBase 分布式数据库

HBase 是一个功能强大的分布式数据存储系统,基于列存储数据记录. 数据行有 3 种基本类型定义:行关键字(Row Key),时间戳(Time Stamp)和列(Column). 每行包括一个可排序的行关键字,是数据行在表中的唯一标示. 一个可选的时间戳,每次数据操作都有一个相关联的时间戳. 某些列中可以有数据也可以没有. 列定义为:〈family〉:〈label〉(〈列族〉:〈标签〉),通过这两部分唯一指定一个数据的存储列. 海量的海洋科学数据经过 MapReduce 计算以后就可以按其 K 值作为行关键字进行分布式存储,实现存储和管理海量数据功能. 海洋有关科学数据的存储如表 1 所示:

表 1 数据存储示例

行关键字	时间戳	列〈ID〉	列〈type〉	
	T_8		type:plant	waterweeds
halobios	T_5		type:anmial	fish
	T_2	1		

对以行名称为 halobios,在时刻 T_2 对列族 ID

的添加数据“1”,在时刻 T_5 对列族 type:plant 添加数据“waterweeds”,在时刻 T_8 对列族 type:anmial 添加数据“fish”.

HBase 主要由主服务器、子表服务器和客户端 3 部分组成. 主服务器作为 HBase 的中心,管理整个集群中的所有子表服务器,监控每个子表服务器的运行情况等. 子表服务器接收来自主服务器的分配的子表、处理客户端的读写请求、缓冲区回收、压缩和分割子表等功能. 客户端主要负责查找用户子表所在的子表服务器地址信息.

平台还可以整合现有的关系型数据库,通过去异构化处理共同提供海量数据存储服务[16-17]. 这里对关系型数据库开发由于篇幅原因不再赘述.

3　海量存储平台特性

以往海洋科学数据存储系统大多采用传统的集群、网格计算技术,耗费昂贵的计算资源且效率不高,可靠性不强等,这里与本平台原型作以下比较如表 2 所示.

表 2　本平台和以往海洋数据存储系统的比较

特性种类	以往海洋数据存储平台	基于 Hadoop 海洋数据存储平台
设计理念	共享数据资源和高性能计算	通用计算和存储平台,共享资源
组成	高端计算机(服务器,集群)	廉价 PC 可实现
功能	单一	丰富,高扩展性,按需增加
性能	低效	高效且高可靠性
容量	可变但有限	按需提供
资源	非虚拟化	虚拟化
应用类型	科学计算	数据处理

综上所述,本平台基于 Hadoop 分布式技术,使编程和实现起来都比较容易,能够高效地存储管理海量数据,具体来说有以下特性:

1) 可扩展性. 具有存储可扩展和计算可扩展性.

2) 经济性. 基于 Hadoop 的海量存储平台可以运行在廉价的 PC 上,无需昂贵的大型机.

3) 安全可靠性. HDFS 的备份恢复机制以及 MapReduce 的任务监控机制保证了分布式处理的可靠性.

4) 高效性. 分布式文件系统的高效数据交互以及本地存储本地计算的处理模式,为高效的处理海量海洋数据作了基础准备.

4　结束语

本文设计并开发了基于 Hadoop 的海量海洋科学数据存储平台. 采用 Linux 集群技术、并行分布式数据库技术、以 Hadoop 分布式平台[14-15]作为基础,主要以 HDFS 分布式文件系统、Map/Reduce 并行计算模型以及 HBase 数据库技术作为处理海量数据方法,在大量的廉价普通计算机上搭建该平台,达到了高效存储和管理北部湾海量海洋科学数据的要求. 目前该海量数据存储平台还在开发中,平台模块实现的结果表明,系统具有良好扩展性和易维护性,系统采用的技术路线和设计方法是有效和可行的.

参　考　文　献

[1] Hayes B. Cloud computing. Communications of the ACM, 2008, 51(7): 9-11

[2] Hadoop. [2010-12-06]. http://hadoop.apache.org/

[3] 陈康, 郑纬民. 云计算: 系统实现与研究现状. 软件学报, 2009, 20(5): 1337-1348

[4] Armbrust M, Fox A, Griffith R, et al. Above the clouds: A berkely view of cloud computing. Berkely, CA, USA: University of California, 2009

[5] Parbhakar Chaganti. Cloud computing with Amazon Web Services. Part 5: Dataset processing in the cloud with SimpleDB. 2009. [2010-12-28]. http://www.ibm.com/developerworks/library/ar-cloudaws5/

[6] Dean J, Ghemawat S. MapReduce: Simplifier date processing on large clusters. Communications of the ACM, 2008, 51(1): 107-113

[7] 李俊, 李勇. 联邦式异构数据库应用系统的集成框架和实现技术的研究. 计算机应用研究, 2001,18(4): 19-22

[8] 科学数据共享工程项目组. 科学数据共享工程门户网站. [2010-12-10]. http://www.sciencedata.cn. 2010

[9] 地球系统科学数据共享网项目组. 地球系统科学数据共享网. [2010-12-10]. http://www.geodata.cn. 2010

[10] 国家测绘局. 测绘科学数据共享服务网. [2010-12-10]. http://sms.webmap.cn. 2010

[11] 国家海洋信息中心. 海洋科学数据共享中心. [2010-12-10]. http://www.mds.sciencedata.cn. 2010

[12] 国家气象局. 中国气象科学数据共享服务网. [2010-12-10]. http://cdc.cma.gov.cn. 2010

[13] 基础科学数据共享服务网. 基础科学数据共享服务网. [2010-12-10]. http://www.nsdc.cn. 2010

[14] 陆嘉恒, 文继荣, 孟小峰, 等. 分布式系统及云计算概论. 北京: 清华大学出版社, 2010
[15] 刘鹏. 云计算. 北京: 电子工业出版社, 2010
[16] Bakis N, Aouad G, Kagioglou M. Towards distributed product data sharing environments. Automation in Construction, 2007, 12(16): 586-595
[17] 余华鸿, 李颖, 张玉川. 数据仓库概述. 计算机与信息技术, 2007, 10(13): 79-99

崔 杰 男, 1986 年生, 硕士研究生, 中国计算机学会会员, 主要研究方向为云计算数据处理、并行分布式计算、信息安全研究.

李陶深 男, 1957 年生, 博士, 教授, 中国计算机学会高级会员, 主要研究方向为分布式数据库、网络信息安全、网络路由算法、无线 Mesh 网络.

兰红星 男, 1956 年生, 博士, 研究员, 主要研究方向为计算机网络与并行分布式计算、网络信息安全.