

基于大数据技术的地震科学数据集成共享研究

刘 坚¹ 马文娟² 李盛乐¹ 谢有顺¹

(1. 中国地震局地震研究所(地震大地测量重点实验室), 湖北武汉 430071; 2. 宁夏回族自治区地震局, 宁夏银川 750001)

摘要: 针对目前地震科学数据共享过程中存在数据种类繁杂、数据量大而存储线性扩展困难、检索性能瓶颈等问题, 从核心地震业务数据资源入手, 梳理了地震观测、探测、实验与试验、专题等 7 大类科学数据; 采用大数据技术, 集成重构了一套全新地震科学数据共享平台。该平台以数据集为基本单位, 为地震科研人员、监测预报人员、地震科学爱好者等全行业用户提供开放、统一、便捷、一站式数据共享服务。

关键词: 地震科学数据; 数据共享; 分布式存储; 大数据; 云计算

中图分类号: P315

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2017.05.010

Integration and Sharing Research on the Earthquake Science Data Based on the Technology of Big Data

LIU Jian¹, MA Wenjuan², LI Shengle¹, XIE Youshun¹

(1. Key Laboratory of Earthquake Geodesy, Institute of Seismology, CEA, Wuhan 430071; 2. Earthquake Administration of Ningxia Hui Autonomous Region, Yinchuan 750001)

Abstract: In the process of sharing seismic scientific data, there are many problems such as complicated data types, large amount of data, difficult linear storage expansion and bottleneck of retrieval performance. This paper starts with the core seismic data resources, and sorts out 7 kinds of scientific data, such as earthquake observation, exploration, experiment and experiment, and special topic. Using big data technology, it integrated a new set of earthquake scientific data sharing platform. The platform takes data sets as basic units and provides open, unified, convenient and one-stop data sharing services for industry users, such as seismological researchers, monitoring and forecasting personnel, and earthquake science enthusiasts.

Keywords: seismological science data, data sharing, distributed storage, big data, cloud computing

1 引言

地震科学数据共享是 2002 年科技部科学数据

据共享首批试点项目之一, 经过 10 多年的建设与发展, 地震科学数据共享工作取得了一系列的成果: 中国地震台网中心率先在地震系统开展地

作者简介: 刘坚 (1978—), 男, 中国地震局地震研究所减灾与遥感应用研究室副主任, 硕士, 研究方向: 地震大数据处理应用研究; 马文娟 (1975—), 女, 宁夏回族自治区地震局高级工程师, 硕士, 研究方向: 地震数据库管理及大数据新技术应用 (通讯作者); 李盛乐 (1965—), 男, 中国地震局地震研究所减灾与遥感应用研究室主任, 硕士生导师, 主要研究方向: 地震分析预报软件研制; 谢有顺 (1992—), 男, 中国地震局地震研究所硕士研究生, 研究方向: 地震信息化研究。

基金项目: 国家科技支撑课题“地震分析预测若干实用技术研究” (2012BAK19B00); 宁夏自然科学基金项目“基于物联网的大数据整合在宁夏地震应急信息交换中的应用研究” (NZ15214)。

收稿时间: 2017 年 7 月 14 日。

震科学数据共享工程, 汇总整理了地震行业众多学科的数据资源, 建设了学科主体数据库, 规划设计了数据管理和服务标准^[1]。随后, 江苏省地震局依托本省的测震数据库、前兆数据库建设了江苏省地震科学数据共享网站^[2], 上海市地震局建设了省局资源共享平台, 对强震、测震、空间信息、地震地质背景等数据进行规范化管理^[3]; 中国地震局工程力学研究所建设了专门针对强震数据共享系统; 中国地震局地球物理研究所上线了测震波形数据共享系统等。然而, 上述共享网站或系统主要采用传统集中式IOE (IBM服务器+ Oracle数据库+EMC存储) 模式, 而随着地震数据爆炸式增长, 该模式下服务器存储量的不断扩展不仅给财政支出和日常运维带来压力, 技术上也存在以下4方面的问题与不足^[4]。一是存储横向线性扩展有限, 传统系统都属于TB级, 而支持PB级能力有限; 二是硬件平台兼容性差, 致使异构平台整合困难; 三是非结构化数据存储处理能力不足, 关系型数据库仅对结构化数据处理高效; 四是数据处理与检索性能瓶颈, 数据量大、读写频繁导致传统数据库I/O性能下降。

针对上述问题, 本文从地震业务核心数据资源入手, 采用大数据技术架构, 将数据采集、清洗、标准化、数据集整理、集成的整个流程, 进行重构设计并实现了一套全新的地震科学数据共享集成平台, 为全行业用户提供开放、统一、一站式数据共享服务。

2 地震科学数据分析

按照地震科学数据分类标准^[5], 地震科学数据分为地震观测数据、地震探测数据、地震调查(考察)数据、地震实验与试验数据、地震专题数据、防震减灾综合数据、其他地震数据等7个大类、44个中类、286个小类。

按时间周期, 地震数据可以分为: 前兆数据、测震数据以及专题整编数据。前兆数据主要是地形变、电磁、地下流体的前兆观测资料, 包括: 原始数据和预处理数据的秒值、分钟值、小时值、日均值, 观测日志和仪器运行日志以及辅

助观测数据(如中国气象局CLDAS大气数据)等。测震数据包括: 2013—2015年SEED、SAC、Miniseed、ASCII等格式的事件波形数据, 1900年至今全球7级以上, 1973年至今全球6级以上, 公元前780年至今全国5级以上, 1965年至今全国2级以上地震目录、全球地震目录、全球灾害信息目录以及中国台网快报、正式观测报告、区域子网观测报告、震源机制解等。专题整编数据:《中国震例》中的285个历史地震资料(1996—2006), 以及倾角 $1^{\circ} \sim 90^{\circ}$ 、滑动角 $-180^{\circ} \sim 180^{\circ}$ 的三维动态断层错动gif图; 大震应急产品数据集(震中位置、地震构造图、地震波传播图、波形图、断层错动模型、震源机制解、地震动预测、余震分布)等。

按时间维度可以将地震数据分为实时、准实时、历史数据三类。

3 地震大数据共享平台总体架构

为实现上述地震科学数据的集成与共享, 克服传统数据共享系统存在的问题, 本文利用大数据技术架构进行地震科学大数据平台的设计, 其总体架构如图1所示。

共享平台主要完成3个层面的工作: 一是采集汇总地震大数据, 二是实现地震大数据的存储处理, 三是地震大数据的展现与查询下载。首先采集汇总三类数据源。其次将汇集的数据置于地震大数据平台进行处理。针对数据的价值高低采取不同融合方式进行存储处理: 事务型业务数据采用关系型数据库存储, 低价值密度数据采用Hadoop下的HDFS (Hadoop Distributed File System)、MapReduce等方式存储处理^[6-9], 高价值密度数据采用列存储、压缩技术、粗粒度索引等技术进行存储管理。最后通过基于台站、测点、测项、类别等方式, 将原始数据、波形、产品、专题等数据以数据集的形式为用户提供查询、预览和下载服务。

地震大数据共享平台, 涉及以下主要关键技术。

(1) 数据标准化。数据标准化包括数据的清

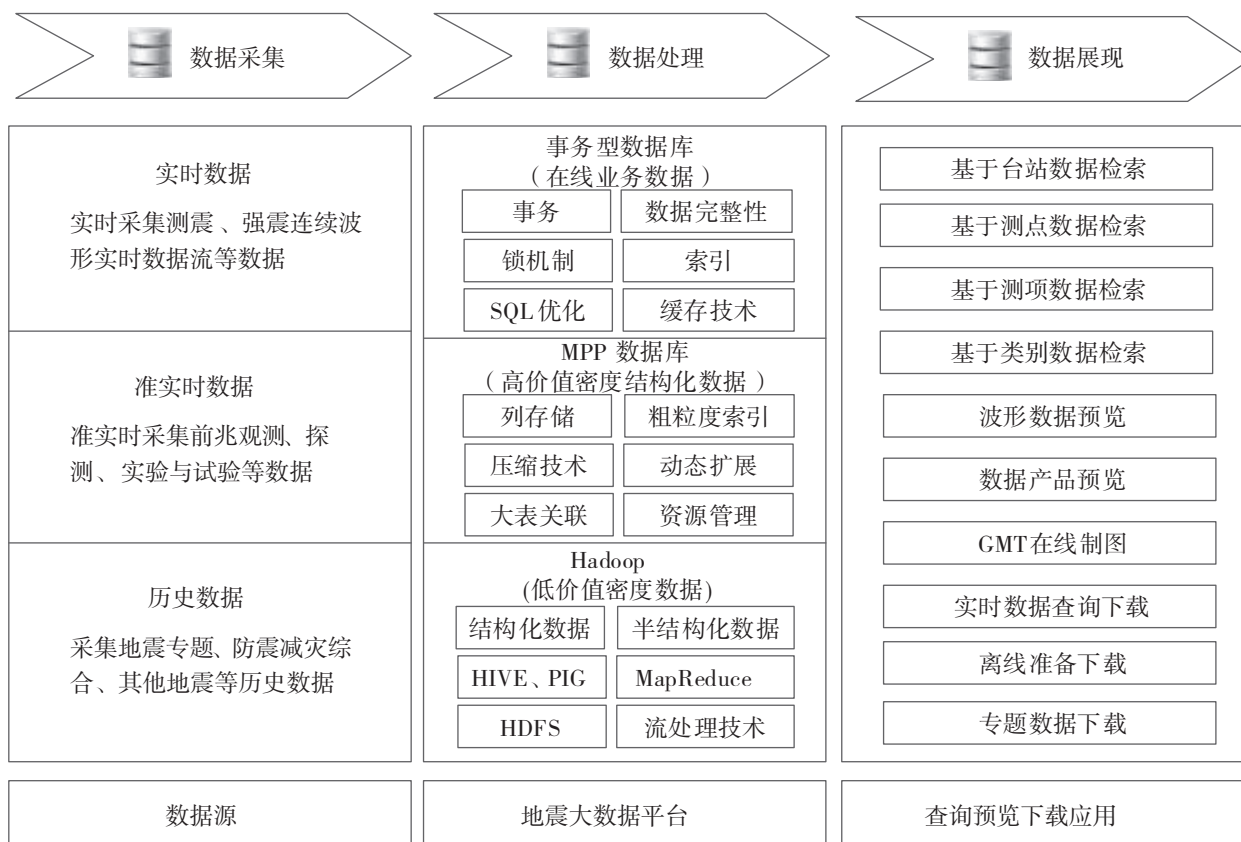


图1 地震大数据共享平台总体架构

清洗处理、数据资源编目和数据格式规范等。其中数据清洗是人工校验与研发数据清洗模型和算法相结合,通过数据间的相互印证和逻辑关系校准异常值,补齐缺失值,形成规范统一的元数据。

(2) 系统互操作。针对跨平台、跨部门、跨系统的不同需求,应通过高通用、可伸缩、易扩展的互操作技术,实现结构化、半结构化和非结构化数据的对接与互连互通。

(3) 数据存储。针对海量数据的实时性、快捷性调用需求,应采用Hadoop和MapReduce等云存储和分布式计算技术,实现存储与计算的并发运行。

(4) 海量数据检索。针对数据产品多源性、海量性和繁杂性特点,规范和统一数据文件管理方式,开发能够支持海量影像数据的快速调用、动态显示的模糊检索和精准检索运算技术,包括布尔检索、逻辑检索、邻近检索和字段限制检索等。

(5) 数据安全。综合利用防火墙技术、安全

评估技术、入侵检测技术、防病毒技术等,建立完整的、立体的、多层次的系统安全防御体系。

4 部分关键技术与设计

4.1 增量数据集成

历史静态数据主要利用Sqoop(在Hadoop和关系数据库中传递数据工具)和Extract-Transform-Load(缩写ETL)工具将数据清洗、标准化后以数据集的形式进行文件存储,处理过程较简单。在此不再赘述^[10]。重点阐述两类增量业务观测数据:一是测震实时波形流数据,二是前兆等准实时数据的集成。从图2可以看到:(1)实时波形流数据处理过程。通过适配器程序将实时波形数据从测震流服务器接入并推送到Kafka(一种高吞吐量的分布式发布订阅消息系统)集群中,相当于一个数据缓冲区,Storm(一种流式处理框架)^[11]上的应用实时获取Kafka消息缓冲区中的数据进行解析、分析,并将结果持久化至HBase和Mysql数据库供共享平台使用和展示。

(2) 前兆和其他准实时数据。利用Sqoop或ETL在监控程序的控制下,将数据源的增量数据抽取至Greenplum(一种在数据仓库中能快速查询结果分布式数据库)供共享平台使用和展示。

4.2 分布式数据存储

针对地震系统内部科学数据量日益暴增使得传统集中垂直方式存在存储扩展瓶颈的问题,可以采用分布式服务器集群水平扩展方法加以解决,而用于集群的服务器可以是普通的服务器,这样既可降低运营成本,也可使存储服务器轻易地扩展至数千台。为此,该平台采用分布式Nosql(Not Only Sql)数据库存储架构设计,地震观测数据如前兆各学科、测震、强震等连续波形、事件波形数据选用Hbase分布式数据库进行存储,而由于HBase具有将所有需要一起进行查询并把数据存储在一起这一特性,HBase集群就自然能够根据key来组织数据。在水平分割时,key值的范围就可以被用来分割数据。每一个服务器可以存储全部数据的一个子集。同时分布式的数据还可以被同时访问。这样大大增强了HBase的可扩展性。

Hbase表是一个分布式多维表,关键在设计好Row Key,以方便数据查询并进行数据分析。根据地震业务逻辑,观测数据表的Row Key可以有以下几个部分构成:<台网Netid><台站Stationid><测点pointid><仪器Intrid><测项Itemid><采样率Samplerate><时间戳Timestamp><产品类别

Protype>。当要查询某个台网某个时间段数据就可以指定起始Row Key为<Netid><Timestamp.MIN_VALUE>,终止Row Key为<Netid><Timestamp.MAX_VALUE>,就可以进行查询了。其他各种组合需求,比如要查询某个自然测点数据、某台仪器的数据、某个学科数据、某个测项分量数据等,皆可非常高效地检索出来^[13]。

通过模拟相同存储环境,将Mysql与Hbase两者针对结构化观测数据的存储进行效能测试,在关键代码行处添加秒表,记录执行命令时间。数据量(条)分别为50条、100条、1000条、1万条、10万条。每次插入保存完毕把所耗时长(单位:ms)写入日志文件。连续多次测试,取平均值。如图3所示,当写入记录条数小于1000时,可以看出两者所耗时间差别并不明显;但随插入记录条数持续增加,Mysql的插入耗时开始剧增,而Hbase耗时变化幅度相对较小,其存储性能优势明显得到提升。

4.3 跨平台异构数据整合

针对共享平台来自不同部门或机构的异构数据,特别是其他部委(如气象局)数据,传统数据整合思路是将数据从各系统抽取过来集中至一个数据中心,进行数据集的物理整合。这种方式不仅涉及重复投资,而且存在数据搬迁困难、整合成本高、非结构化数据整合困难等问题。为此提出一种新的整合方式:基于分布式云计算服务的逻辑数据整合。该方式的整合思路为:不强

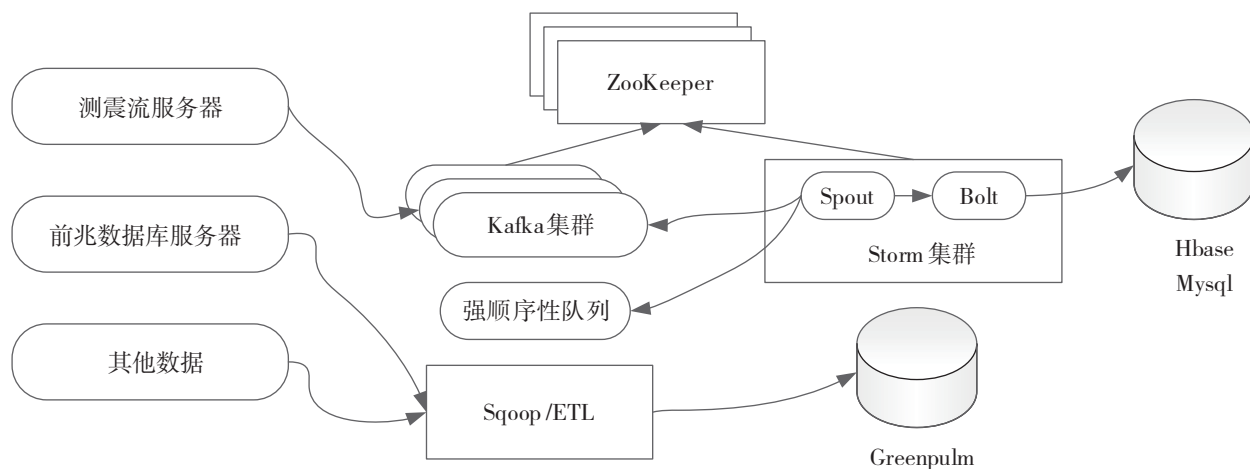


图2 增量数据集成示意图

求物理上的集中,而是保持原部门或机构数据的分布现状,将各个系统的数据通过接口包装成服务,如图4所示,注册到企业服务总线,通过企业服务总线提供统一的数据服务,从而实现数据在逻辑上的整合。

(1) 结构化数据的整合过程

作为数据源的结构化数据库需要开放数据库接口,供元数据管理系统从源数据库中抽取数据结构信息,并保存在元系统中。服务生成模块

可以查询存放于元数据系统中的各业务系统元数据,通过简单的操作自动生成提取数据的代码块,并将该部分代码块包装成云服务,存放于服务运行模块中,并服务注册到企业服务总线,对外部进行数据服务。

(2) 非结构化数据的整合过程

对于NoSQL数据库,由于没有统一的数据结构,是无法通过上面的方式自动生成代码块并发布提供服务的。但可以通过定制服务接口的方

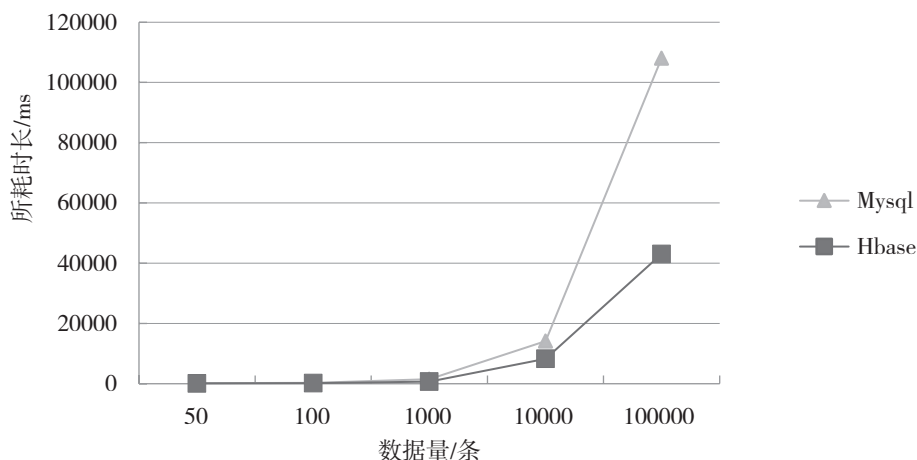


图3 与关系数据库存储性能对比

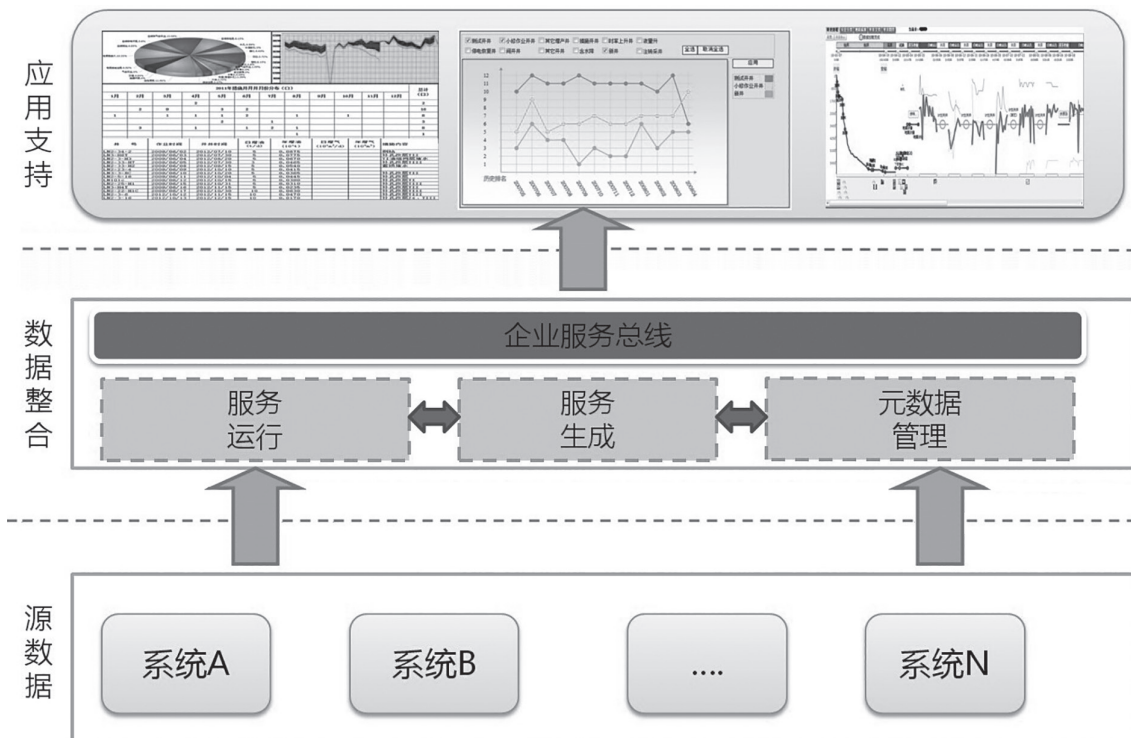


图4 异构数据整合示意图

式生成提供服务，通过云服务进行集成并发布到数据整合平台，统一对外提供服务。在这种情况下，只能针对每个接口进行云服务的定制开发。

4.4 数据格式标准化

平台数据种类多，涉及众多学科，规范不统一，具有典型多维、异构特点，且用户需求也各异多样，为此平台采用“化整为零，各个击破，分而治之”策略，对于震源机制解、地震目录等已规范化的测震数据，保留原国际通用的标准格式，亦便于用户直接下载使用；对于测震连续、事件波形数据，因数据量大，结构复杂，则采用MiniSEED^[12]格式，既满足专业用户需求，又减少数据格式转换代价；对于涉及多学科的前兆数据，虽然数据种类各异，但因其业务参数基本相同，平台利用XML封装设计了一种通用的、易于理解、易于解析的中间数据格式模型。

数据集XML数据封装模型，结构如下：

```
<?xml version=" 1.0" encoding=" utf-8" ?>
<Dzdataset>
  <metadata>
    <!-- station: 台站代码; stationpoint: 测点
    代码; stationitem: 测项分量代码; sample: 采样率
    代码; cptype: 产品编码; computetype: 计算方法;
    dinstr_finstr: 其他测点编码 --->
    <station description=" 台 站 代 码">14001</
    station>
    <stationpoint description=" 测 点 代 码">3</
    stationpoint>
    <stationitem description=" 测 项 分 量 代 码
    ">2212</stationitem>
    <sample description=" 采 样 率 代 码">60</
    sample>
    <cptype description=" 产 品 编 码">0</cptype>
    <computetype description=" 计 算 方 法">0</
    computetype>
    <dinstr_finstr description=" 其 他 测 点 编 码
    ">00</dinstr_finstr>
  </metadata>
  <datasets>
```

```
<!-- starttime: 起始日期; pd: 兼容考虑参
数; processingflag: 审核标记; defaultnum: 缺数个数;
excepttype: 异常类型; exceptgoal: 异常指标;
combine: 其它综合字段 --- description=" 其他测
点编码">
```

```
<set>
  <set-description>
    <samptime description=" 采 样 时 间
    ">2017-06-08 00:00:00.0</samptime>
    <pd description=" 兼 容 考 虑 参 数
    ">null</pd>
    <processingflag description=" 审 核 标 记
    ">0</processingflag>
    <defaultnum description=" 缺 数 个 数
    ">0</defaultnum>
    <excepttype description=" 异 常 类 型
    ">0</excepttype>
    <exceptgoal description=" 异 常 指 标
    ">0</exceptgoal>
    <combine description=" 其 他 综 合 字 段
    ">0</combine>
  </set-description>
  <data>-788.99 -791.35 -793.16 -792.17
  -789.12 -785.62 -781.13 -777.04 -776.08
  -777.42 -779.74 -784.42 -790.77 -795.88
  -799.42 -800.99 -800.6 -797.11 -793.67 -787.58
  -783.51 -782.27 -782.7 -784.29
  </data>
</set>
</datasets>
</Dzdataset>
```

XML数据封装模型的定义基于地震前兆业务逻辑为基本结构框架，以标签Dzdataset为根元素标识数据类型，主体部分由元信息部分和数据集部分组成，分别以标签metadata和datasets标记。

平台设计的XML文档依据前兆数据库管理规则，以统一的数据模型来描述保存在异构环境中的各种数据，灵活表达数据内涵，屏蔽了数据

源中应用环境和数据结构的异构性。同时,标签元素的规整性也确保了文档在网络中传输的便利性以及平台处理数据的统一性^[14]。这些特征有利于XML文档在不同系统之间的交换,推动数据的全面共享。

5 共享平台实现

平台实现主要是服务器端网站研发,为满足平台“低耦合、高内聚”的目标,实现代码的健壮性和可扩展性,平台采用典型的三层结构,即用户界面层/表示层(负责与用户进行交互)、业务逻辑层、数据访问层(主要是对非原始数据的操作层,也就是仅对数据库,而不是对数据的操作,具体为业务逻辑层或用户界面层提供数据服务)。本平台采用MVC模式的SSM

(SpringMVC、Spring和Mybatis)框架实现,如图5,各层采用不同的技术构建了易于维护扩展的Web应用程序。

地震数据共享服务平台以友好的门户网站向地震全行业用户开放,主页链接为:<http://10.5.109.26:8080/csds/index.html>。如图6所示,客户端用户仅需在浏览器地址栏键入主页链接即可进入系统主页。主页包括5个部分:导航栏、重要资料推荐区、图形展示区、快捷搜索区和产品滚动区。

系统上线以来,运行稳定,具有可靠的容错机制;支持多用户并发操作;系统单次响应最大时间 ≤ 20 秒;海量数据平均单次查询 ≤ 30 秒;离线数据准备最大时间 ≤ 24 时;在服务端可动态添加业务service模块,提供接口使用,在功能和

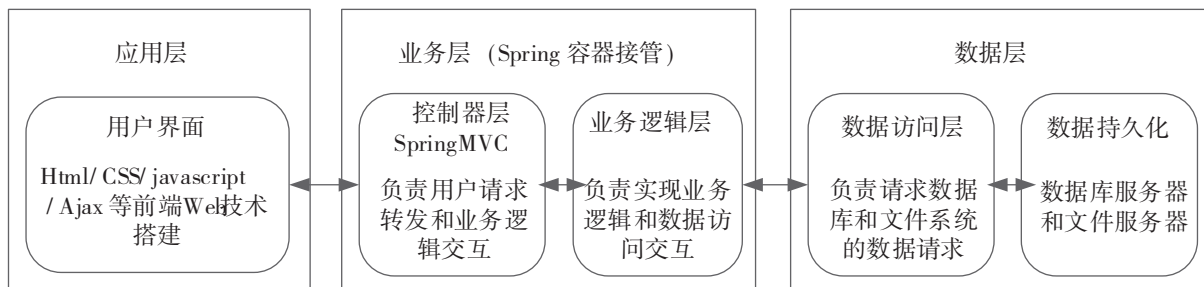


图5 服务器端开发技术框架示意图



图6 平台主页

性能上均达到预期要求。

6 结论与讨论

(1) 采用分布式数据库(Nosql)技术,解决了传统集中式存储水平扩展困难的问题,与传统关系型数据库相比,数据量越大,Nosql数据库存储与检索的效能优势越明显。

(2) 采用云计算技术,解决了数据搬迁困难、整合成本高、非结构化数据整合困难等问题,使不同部门或机构的异构数据源整合变得快捷高效。

(3) 采用了数据标准化技术,利用XML数据封装模型实现了异构环境下数据的自由、便捷共享,提供了涵盖所有数据结构的多维度检索及相应的下载服务,从不同角度满足了用户的检索、下载需求。

利用上述大数据技术,地震科学数据集成共享平台重构建成后,为地震全行业用户提供了无障碍、一站式数据共享渠道,全面提升了行业数据源的共享服务能力,但随着共享平台的推广使用,主体共享数据和用户量都将不断扩充,数据的分布式管理和在线可视化功能需进一步完善,以保障系统的高效、稳定和体验度。

参考文献

- [1] 刘瑞丰,蔡晋安,彭克银,等.地震科学数据共享工程[J].地震,2007,27(2): 9-16.
- [2] 詹小艳,许红梅,朱升初,等.江苏省地震科学数据共享平台技术研究[J].防灾科技学院学报,2012,14(1): 57-63.
- [3] 袁媛,尹京苑.上海地震科学数据共享服务平台建设[J].华北地震科学,2014,32(1): 16-21.
- [4] 诸云强,宋佳,潘鹏,等.地学数据共享发展现状、问题与对策研究[J].中国科技资源导刊,2014(4): 55-63. DOI: 10.3772/j.issn.1674-1544.2014.04.010
- [5] 黄永文,张建勇,黄金霞,等.国外开放科学数据研究综述[J].现代图书情报技术,2013(5): 21-27.
- [6] LIVNY M, THAIN D, TANNENBAUM T. Distributed computing in practice: the condor experience[J]. Concurrency & Computation Practice & Experience, 2005, 17(2/4): 323-356.
- [7] AGRAWAL D, DAS S, EL A A. Big data and cloud computing: current state and future opportunities[C]//14th Int'l Conf. on Extending Database Technology (EDBT 2011). Uppsala: ACM Press, 2011: 530-533. DOI: 10.1145/1951365.1951432.
- [8] TALLON PP. Corporate governance of big data: perspectives on value, risk, and cost[J]. Computer, 2013, 46(6): 32-38. DOI: 10.1109/MC.2013.155.
- [9] TALIA D. Clouds for scalable big data analytics[J]. Computer, 2013, 46(5): 98-101. DOI: 10.1109/MC.2013.162.
- [10] 孙大为,张广艳,郑纬民.大数据流式计算:关键技术及系统实例[J].软件学报,2014,25(4): 839-862.
- [11] Storm wiki[EB/OL].[2014-09-15]. <http://en.wikipedia.org/wiki/Storm>.
- [12] IRIS.Data Formats[EB/OL]. [2016-05-18]. <http://ds.iris.edu/ds/nodes/dmc/data/formats/#miniseed>.
- [13] 刘坚,李盛乐,戴苗,等.基于Hbase的地震大数据存储研究[J].大地测量与地球动力学,2015,35(5): 890-893.
- [14] 唐晓光.基于订阅机制的数据共享平台的研究与设计[D].大庆:东北石油大学,2013.

(上接第73页)

- [5] CASTERS M, BOUMAN R, DONGEN J V. Pentaho Kettle解决方案:使用PDI构建开源ETL解决方案[M].北京:电子工业出版社,2014.
- [6] 郭丹,樊红.基于ETL-KETTLE的贵州卷烟营销大数据分析可视化[J].计算机系统应用,2017,26(1): 74-80.
- [7] 曹一化,刘旭.自然资源资源共性描述规范[M].北京:中国科学技术出版社,2006.
- [8] 王运红,张莞,沈欣媛,等.国家自然资源资源e-平台建设实践[J].中国科技资源导刊,2008(4): 16-19. DOI: 10.3772/j.issn.1674-1544.2008.04.004.
- [9] 刘龙庚,杨东日,李小平.信息产业资源共享与分析平台中ETL技术研究[J].四川大学学报(自然科学版),2012,49(1): 85-89.