

ETL在科学标本数据规范化集成整编中的应用

吴志远 杨眉 何明跃 施爽
(中国地质大学(北京), 北京 100083)

摘要: 在科学标本数据的规范化集成整编中, 标准和规范是前提, 资源数据的整合、集成是核心。ETL技术能够实现数据的抽取、转换和装载。本文将其引入到科学标本数据规范化集成整编中, 实现对多源的、异构的标本数据的汇集和整合, 并以岩矿化石标本数据集成整编为例进行应用实践, 取得了较好的应用效果, 证明了应用ETL技术实现标本数据规范化集成整编的有效性。该方法具有一定的实践借鉴意义, 可为其他资源数据的规范化集成整编提供一定的科学参考。

关键词: ETL; 科学标本; 数据集成整编; 数据整合; Kettle

中图分类号: TP311.13

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2017.05.009

Application of ETL in Standardized and Integrated Reorganization of Scientific Specimen Resource Data

WU Zhiyuan, YANG Mei, HE Mingyue, SHI Shuang
(China University of Geosciences, Beijing 100083)

Abstract: In the standard and integrated reorganization of scientific specimen resource data, standards and norms are the prerequisite, and the collection and integration of resource data are the core. ETL technology has the ability to realize data extracting, transforming and loading. this paper introduces into the standard and integrated reorganization of scientific specimen resource data to collect and integrate multi-source and heterogeneous specimen resource data, and proves the effectiveness of ETL technology in the standard and integrated reorganization of specimen resource data. The method has a certain practical reference significance, providing some scientific reference for the standard and integrated reorganization of other resources data.

Keywords: ETL, scientific specimen, integrated reorganization of data, data integration, Kettle

1 引言

随着“大数据”时代的来临, 资源单位对于馆藏标本的数字化、信息化也在迅速推进, 科学标本数据出现快速增长的趋势。虽然我国已制

定了八大领域标本数据规范, 但如何通过技术手段实现科学标本数据的高效集成整编, 形成规范化、系统化的数据, 并达到长期、持续的整合集成, 是数据整编工作中的技术难点。ETL是解决构建数据仓库, 实现数据整合集成的主要技术。

作者简介: 吴志远 (1979—), 男, 中国地质大学(北京)工程师, 研究方向: 数据库技术(通讯作者); 杨眉 (1980—), 女, 中国地质大学(北京)助理研究员, 研究方向: 标本资源数字化与共享; 何明跃 (1963—), 男, 中国地质大学(北京)教授, 研究方向: 矿物学岩石学矿床学; 施爽 (1992—), 中国地质大学(北京)硕士研究生, 研究方向: 矿物学。

基金项目: 科技基础性工作专项重点项目“科技基础性工作数据资料集成与规范化整编”(2013FY110900); 国家科技基础条件平台项目“国家岩矿化石标本资源共享平台”。

收稿时间: 2017年7月14日。

ETL能解决科学标本数据整合面临的诸多问题,实现海量、异构标本数据的高效、持续整合,具有很好的适用性以及性能方面的优势。本文应用ETL技术有效整合标本信息资源,实现科学标本数据的规范化集成整编,对于建立科学标本资源数据库,实现标本资源的信息共享和有效利用具有重要意义。

2 ETL 技术

ETL^[1]即数据抽取(Extract)、转换(Transform)、装载(Load)的缩写,代表从数据抽取到装载的技术实现过程,是一个用来实现异构多数据源的数据整合集成工具。ETL技术可以实现数据从数据源向目标数据仓库的转化,此转化是实施资源整合的重要步骤^[2]。ETL主要包括3个环节:一是抽取,将数据从分散的异构数据源中读取出来,这是所有工作的前提;二是转换,按照预先设计好的转换规则对抽取的数据进行转换,使本来异构的数据转换后能达到统一的标准;三是装载,将转换好的数据按照计划增量或全量加载到数据仓库中^[3-4]。ETL处理流程是:首先用一个验证步骤去确定到达的或被抽取的数据是哪一种类型,然后数据被送到一个特定转换进行处理。当转换完成后,数据将被传递到下一个转换或者一个目标表中,并在发生错误的情况下,被转移到一个错误流程进行处理^[5]。“ETL”概念结构如图1所示。

Kettle^[6](Kettle E.T.T.L. Environment)是一款国外开源的ETL工具,纯Java编写,可在Window、Linux、Unix上运行,绿色版无需安装,数据抽取高效稳定^[7],在国内数据转换、处理中应用较多。它允许管理来自不同数据接口的数据,通过提供一个图形化的用户环境来描述你想做什么,而不是你想怎么做。Kettle支持很多种输入和输出格式,包括文本文件、数据表以及商业和免费的数据库引擎。

3 标本数据整编方案

数据整编的关键是按照统一的技术标准,通

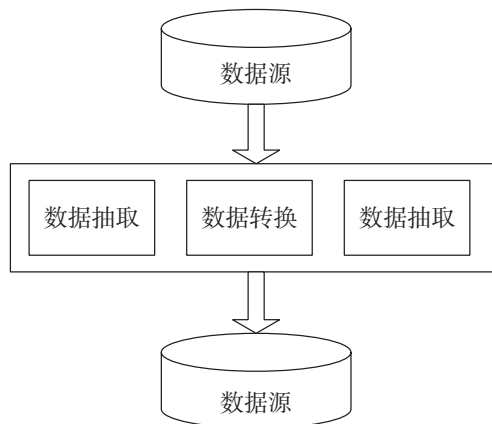


图1 “ETL”概念结构图

过数据转换处理,打破项目边界,将多源(可能是不同的基准或数值单位)不同空间、时间,相同要素的数据集成在一起。首先需要重点确定标本数据要素对象、要素属性全集,统一属性项语义标准、值域范围及数值单位。在此基础上,制定标本数据资料整编方案,内容包括:整编技术标准、数据库表结构、整编软件工具、整编实施等。

3.1 整编技术标准

要进行标本数据集成整编,标准和规范必须先行^[8]。根据科技基础性工作专项制定的相关数据标准与规范,标本数据整编参照的技术标准主要有:科技基础性工作专项项目科学数据汇交标本资源描述规范、科技基础性工作专项数据资料分类与编码。

3.2 数据库设计

数据库设计一般有需求分析、概念模型设计、逻辑模型设计、物理设计、测试修改和数据字典编写6个阶段。标本数据库设计时,参考这6个阶段,主要依据现有的整编技术标准,分析要素对象及属性字段,设计标准化的要素数据表结构。对于岩矿化石标本,根据自然科技资源共性描述规范,其要素数据表的主要属性项包括:平台资源号、资源编号、资源名称、资源外文名、保存资源类型、保存资源数量、共享方式、获取途径、保存单位、资源归类编码、具体用途、简要特征描述、资源提供者、资源提供时间、经度、纬度、高程、产地、省、国家、地质

产状或层位、资源形成时代、标本编号、采集号、库存位置号、实物状态、图片名、主要用途。此外,增加“标本资源唯一编号”“对应的元数据编号”“所属项目编号”3个字段。其中,所属项目编号、对应的元数据编号与汇交的数据文件相同,标本资源唯一编号形式是:专项数据资料分类编码+流水号。除要素数据表外,还需要设计相关的数据字典。

3.3 整编软件工具

通常技术上进行数据整合、集成的方法有:手工编写代码和利用软件工具。采用手工编写代码的方式,开发效率低、周期长、难度大,容易出错,缺乏一致性的日志和错误处理,难以保证可维护性,进行大数据量的数据整合时可能会遇到性能问题。ETL技术在对数据源的支持、数据转换、管理和调度、集成和开放性、对元数据管理方面都有很好的表现,采用ETL工具作为整编软件工具可以减少代码工作量,能够较好地满足数据整编中数据多源、异构及整合效率、性能的要求。

3.4 整编实施

对原始标本资源数据的数据来源、要素及属性、内容格式等进行分析,并根据整编技术标准和数据库设计,在ETL工具中为各种数据源设计数据转换作业流程。通过调度运行ETL作业,进行数据抽取、合并与集成等操作,将标本资源数据按照统一的标准整合到目标数据库中,为后续标本资源的数据管理、共享、挖掘、分析等应用提供基础数据源。

4 岩矿化石标本数据的集成整编

岩矿化石标本资源是自然资源领域八大类^[9]资源之一。目前,对于岩矿化石标本资源的数据整编,采用的是分布式加工方式,即各个资源单位向数据中心汇交标本数据或提供数据接口,然后由数据中心统一对各数据源进行规范化集成整编。

依照科学标本数据整编方案,整编的主要流程是:首先参照岩矿化石标本资源相关的标准规

范设计数据库,选用大型关系数据库Oracle;然后采用ETL工具Kettle作为主要整编工具软件,在Kettle中设计数据整合流程;最后调度执行任务,实施数据整编。其中的应用ETL技术实施数据整编是整个过程的关键,也是本文分析研究的重点。使用kettle进行岩矿化石标本资源数据规范化集成整编过程如图2所示。

4.1 数据库设计

参照“科技基础性工作专项项目科学数据汇交标本资源描述规范”中的岩矿化石资源描述规范表设计岩矿化石标本数据库及核心数据表的数据字典。从逻辑结构进行划分,岩矿化石标本数据库的主要构成有表格数据和图片数据。

表格数据:用二维表格表示的数据,包括岩矿化石标本数据表、岩矿化石分级归类与编码表和资源单位表。其中矿化石标本数据表是数据库主表,根据数据字典设计,共29个字段;岩矿化石分级归类与编码表的设计参照“科技基础性工作专项数据资料分类与编码”中的岩矿化石编码部分,表结构采用目录树的形式。标本数据表中的资源归类编码字段外键关联编码表的主键。

图片数据:拍摄的岩矿化石标本图片文件,文件名保存在标本数据表图像字段中。

4.2 数据源准备

数据整合之前,首先要对岩矿化石标本数据源进行调研分析,了解各单位数据源的数据格式、内容、采集手段、数据质量等。由于标本资源来源于不同的单位,其数据是离散、异构的,关联程度较低,并且各资源单位汇交数据的方式也不尽相同,主要有以下几种方式。

一是离线文件方式,通常是Excel、Access、Txt等数据文件。

二是数据库连接方式。通过数据库连接字符串连接Oracle、Mysql、Sqlserver等关系数据库,这种方式由于对外暴露了数据库连接参数,存在安全隐患,只能是临时或内部使用。

三是在线接口方式。资源单位通过HTTP对外提供接口服务,目前比较流行的有Rest、Soap接口,使用接口方式连接方便,但对于技术水

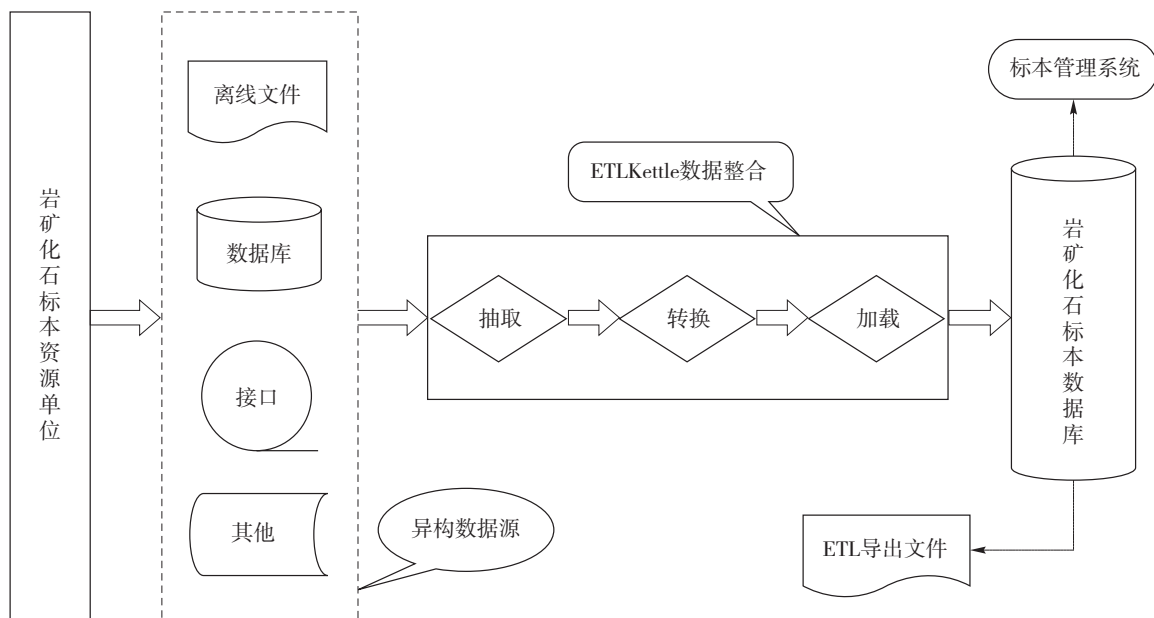


图2 岩矿化石标本 Kettle 数据整编过程

平要求比较高,需要资源单位能开发出相关 Webservice 接口并对外提供服务。

Kettle 对这几种常用的数据源都有很好的支持作用,其内置有多种输入组件支持在不同类型数据源之间建立连接,屏蔽各种数据源之间的差异,为整合输入提供统一的数据视图。

4.3 ETL kettle 数据整合

利用 ETL 工具 Kettle 设计整合流程。Kettle 提供了一个图形用户界面 Spoon。Spoon 创建了转换和作业的数据集成环境,可以通过 Spoon 快速地设计转换和作业。转换和作业可以保存在 Kettle 资源库中,便于以后转换和作业的复用。从资源库中加载作业与转换流程,可以方便地实现对作业、转换流程的优化与重新定制,提升 ETL 处理效率,改善数据质量。根据 Kettle 的功能特点,设计如下整合流程。

(1) 数据抽取

数据抽取是从不同方式的数据源如文件、数据库、接口中抽取数据。按照抽取方式的不同,需要在 Kettle 中为每一种数据源选取相应的内置输入组件,设计单独的整合流程。按照抽取时间的不同,对数据 ETL 过程可以分为两种类型:全量 ETL 过程和增量 ETL 过程。全量 ETL 过程一般用于数据库的初始化,而增量 ETL 过程则用于

数据库的后期增量维护。设计 Kettle 流程时可根据需要选择全量抽取或增量抽取,一般采取首次全量、以后增量的方式。由于整编标准规定了每条岩矿化石标本数据具有一个唯一的资源编号,因此在增量数据抽取时,可根据目标数据库该资源编号存在与否由 Kettle 自动选择进行数据覆盖更新或增量添加。

(2) 数据转换

数据转换首先需建立数据映射。数据的映射指的是建立从源数据到目标数据的映射关系。实际整合中,部分单位的标本数据是严格按照整编标准录入的,转换过程建立映射关系即可,无需进行其他转换;但还有许多单位的标本数据,与整编标准存在一定的差别。一些来自于单位原有数据库或数据文件,其数据字段、内容、格式等不符合标准规范,甚至某些字段在数据标准中并不直接存在,而是需要根据某些算法公式或者某些计算公式对各部分数据进行运算才能得到。对于这些数据必须通过转换来实现数据标准的统一。主要包括以下几种数据转换方式。

数据格式转换:将来自不同数据源的同类数据转换为相同格式。

数据类型转换:将某种数据类型或格式转换为另一种数据类型或格式的较低层次转换^[10]。

数据选择操作：根据选择条件对数据源中的数据进行选择。

字段抽取：从一个或多个数据源中抽取有用的字段。

字段合并：通过计算、导出、分配，从多个数据源中将需要的字段整合在一起。运用在需求处理过程中确定的业务法则进行各种转换，包括字符串操作、日期和时间算术运算、条件语句以及基本计算等功能。

聚集和概括：将某一实体的实例数目减少到易于进行直接利用的水平，也有助于预先计算出广泛应用的概括数字以使每个查询不必重新计算。

例如，针对岩矿化石标本资源单位提供的Rest接口和数据库连接方式的数据源，在Kettle中设计的数据转换流程如图3所示。

在图3中，上面一行是Rest接口数据源的整合流程，通过Kettle Rest Client组件将Rest接口地址输入转换成Json数据输入，经过一系列字段选择、字符串操作、字符串替换转换操作，以插入更新的方式（资源编号不存在即插入，存在则更新）将标本数据整合到岩矿化石数据库中。下面一行是数据库连接方式的数据源，通过连接资源单位的岩矿化石数据库表抽取数据，再经过中间相关的转换操作，最后同样以插入更新的方式

整合数据到目标数据库。将前面定义好的转换流程整合成一个ETL过程作业，如图4所示。

(3) 数据加载

数据在完成数据抽取、转换后，按照统一的数据格式进行存储和加载。加载后的目标数据库完全满足整编技术标准，从而实现了从多源异构数据到统一目标数据库的整合集成。目标数据库为岩矿化石标本资源管理系统提供标准的数据源，还可以作为Kettle转换的数据源，导出生成其他数据格式如Excel文件。

4.4 ETL kettle作业调度

在准备好数据源、目标数据库，并设计完成Kettle转换、作业流程后，进行最后一步的任务执行。Kettle任务执行也就是实现数据加载的过程，一般是通过作业调度来实现的。作业是对转换过程的整合，Kettle任务执行作业调度有以下几种方式。

(1) 图形界面执行作业。在Kettle Spoon集成环境点击运行作业，这种方式需要手工执行，进行作业调试时使用较多。

(2) 命令行执行作业。通过编写Shell脚本调用Kettle作业命令运行程序Kitchen。Shell脚本可手工执行，也可以在Linux下添加Crontab任务来实现作业的定时调度。

(3) API调用执行作业。通过Java程序调用

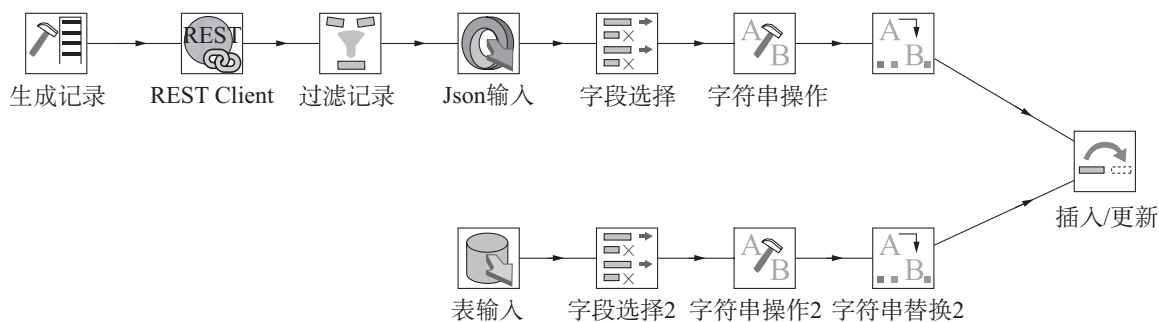


图3 Kettle数据转换流程

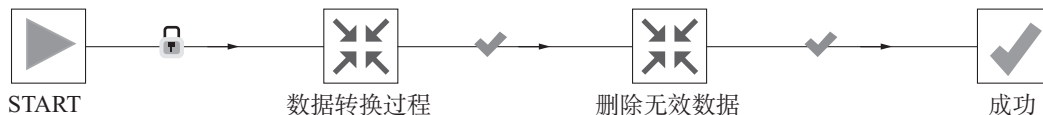


图4 Kettle作业过程

Kettle 提供的 JAVA API 运行任务。可以灵活地通过程序自行定制、批量处理作业,技术上要求较高,与 Java 开源组件 Quartz 配合可以实现程序中作业的定时调度。

标本数据整编实施中,应根据数据源选择 Kettle 作业调度方式。离线文件方式的数据源,一般不需要定时调度,可通过图形界面执行或 Shell 脚本手工执行;数据库连接和接口方式的数据源为在线方式,采用定时调度的方式,无需人工干预,可根据情况使用 Shell 脚本或 Java 程序来定时执行作业。整编后的目标数据库由岩矿化石标本资源管理系统负责后台数据管理,为了便于对多个整合流程的 Kettle 作业进行管理,通过系统后台程序调用 Kettle 提供的 JAVA API,将 Kettle 的作业调度集成到资源管理系统的功能中。在系统管理页面中可以很方便地手工调度 Kettle 作业,还可以结合 Quartz 设置定时任务,由系统后台定时执行 Kettle 作业。

4.5 应用效果

依照制定的整编方案,通过对岩矿化石标本各资源单位的数据源进行分析、利用 Kettle 设计数据整合流程并调度运行作业,成功应用 ETL 技术实现了标本资源的规范化集成整编。目前已成功整合的岩矿化石标本数据有 10 多万条。对于异构数据源,在设计好流程后,整编输入数据量越大,整合效率越高。经测试,对于 1000 条数据输入, Kettle 作业运行时间一般不超过 10 秒;对于 1 万条左右的数据,一般不超过 30 秒。与手工编程相比,极大地提高了整合速度。

采用保存 Kettle 整合流程在资源库中的方式,在后续新增的数据源在数据格式没有变化的情况下,可以直接复用流程;当发生变化时,可根据数据源变化对流程进行相应的调整,无需重新设计流程,提高了整合效率。通过调用 JAVA API 将 Kettle 作业调度集成到标本资源管理系统中,实现了整合流程的作业调度系统化、自动化管理。

在整合过程中,一些因素会导致作业的调度运行出现异常。如数据源的质量问题(数据格式、长度、一致性等),经常会引起作业的异常终止。

通过分析 Kettle 的运行日志中的错误信息,确定数据源中具体哪一部分的数据存在问题,然后将相关问题反馈给资源所在单位,由资源单位负责进行修改。因此, Kettle 作业运行时可以校验数据的完整性、一致性等要素,保证数据质量。当数据源质量较差时,会严重影响整合效率。

5 结语

(1) 在标本数据规范化集成整编中,由于标本数据的多源异构性,导致了整编工作的复杂性。ETL 技术能够高效地将异构数据源中的标本资源数据进行抽取、转换和加载,很好地满足了规范化集成整编中数据整合的需求。

(2) 根据需求设计了标本资源数据整编方案,以岩矿化石标本资源数据为例,使用开源 ETL 工具 Kettle 设计实现了异构标本数据源的规范化数据整合、集成过程,较好地解决了规范化集成整编中由于数据多源异构性而造成的数据难以实现集成的问题。

(3) 由于 ETL 是按照统一的标准和规范进行数据整合,对于源数据中不符合标准规范并且无法进行数据转换的私有数据部分(资源单位自定义),ETL 将按照规则进行过滤。与源数据相比,整合集成的岩矿化石标本目标数据会损失这部分私有数据内容。

(4) 由于 ETL 技术的复杂性和标本资源数据的多源异构性,探索一套通用的 ETL 数据整合方法,使其更加自动、高效和智能,有待下一步研究。

参考文献

- [1] 廖嘉嘉,邓苏,刘青宝.ETL 综述[J].计算机工程,2004,30(3): 4-5.
- [2] 高立春,徐叶强.ETL 在公安部门数据共享与资源整合中的应用[J].情报杂志,2011,29(B12): 190-192.
- [3] 尹晓楠,邹晓涛,张冬.基于 Kettle 的北京市水务普查数据的提取与转换[J].中国水,2013,21: 57-59, 42.
- [4] 崔有文,周金海.基于 Kettle 的数据集成研究[J].计算机技术与发展,2015(4): 153-157.

(下转第 81 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.