基于 Duckling 应用集成框架的数据集成与共享工具 DLM①

李方鑫 南 凯 于建军 (中国科学院计算机网络信息中心 北京 100190)

摘 要: 虚拟科研应用环境中多源异构数据的集成和共享问题是个热点问题。本文基于虚拟科研环境协同工作套件 duckling 应用集成框架，实现了一个通用的数据集成与共享工具 DLM。该工具使用 Portlet 技术，是 BS 架构的 Web 应用，提供通用的数据采集、数据处理、数据可视化展示和数据下载功能，解决了虚拟科研环境中数据的统一采集、处理和共享问题。该工具已经在大气、气象和生物等相关领域中的课题研究中得到了应用，取得了初步的成效。

关键词: 信息集成工具;多源异构数据;数据集成;数据共享;数据可视化展示

Data Lumber Mill Based on Duckling Application Integration Framework

LI Fang-Xin, NAN Kai, YU Jian-Jun

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Integration and visualization for heterogeneous scientific data is a hot topic in e-Science applications. This paper presents generic information integration and sharing tool based on virtual research environment workbench (duckling) application integration framework. The tool uses portlet technology, is a web application, proving generic data collection, data process, data visualization and data download function, solves the data collection, process and sharing problem in s-science applications. This tool has been realized and used in the atmosphere, meteorological and biological- related projects.

Keywords: information integration tool; heterogeneous data; data integration; data sharing; scientific data visualization

1 引言

当代观测型科研活动中会产生大量的科学数据。这些科研活动以及产生的数据一般具有如下特点：

科研活动中产生的数据一般需要进行采集、处理、展示和共享，这是科研活动顺利进行的前提，是当代科研活动的必要环节；科研活动中会有大量的数据采集点(简称站点)，这些站点一般是分散的；科研活动经常需要多个单位联合进行，产生的数据的数据格式经常会多种多样。

科研活动中会有许多应用，虽然每个应用拥有独有的数据源和数据格式，但是它们在数据采集、处理和共享的方式上有很多相同点，拥有相同的数据处理流程和模型。目前，虚拟科研环境中还没有通用的数据集成和共享工具，这样的项目只能重复建设。

针对以上问题，在分析了数据采集、处理和共享的需求，并经过业务抽象和设计后，本文设计并实现了一种基于 Duckling 应用集成框架的数据集成与共享工具 DLM(Data Lumber Mill)，形成了通用的数据集成和共享流程，实现了数据的统一采集、处理和共享。该工具能够有效满足科研活动在数据集成和共享

① 基金项目:e-Science 虚拟科研平台研究与开发(INFO-115-D01);科学数据网格及科研应用系统(2006AA01A120);面向蛋白质科学的高性能计算研究(KGGXI-YW-13)

收稿时间:2010-03-12;收到修改稿时间:2010-04-01

方面对通用性和定制性的需求，避免了数据集成和共享流程的重复建设，有效支持了科研活动。

2　DLM介绍

2.1 DLM 框架

2.1.1 Duckling 介绍

支持 e-Science 的协同工作环境套件 Duckling 是由中国科学院计算机网络信息中心协同工作环境研究中心开发，专为科研团队提供的综合性资源共享和协作平台[1]。经过数年的不断发展，现在的最新版在 Portlet JSR286 规范下融合了文档协同工具(DCT)、文档库管理工具(CLB)、以及用户管理工具(UMT)，提供良好的国际化支持。同时提供了一种机制，可以在 Duckling 门户快速集成自定义开发的其他应用。

2.1.2 Duckling 应用集成框架介绍

Duckling 环境可以为用户提供协同编辑，文档管理，用户管理和授权等功能。Duckling 应用集成框架集成了一个 Portlet 标准的实现 Pluto 2.0。实际应用中的特定问题，可以按照应用插件的方式集成到 Duckling 系统中，从而为实际应用提供了定制化的功能和一个相对完整的协同科研环境。下图 1 为基于插件的 duckling 应用集成框架。

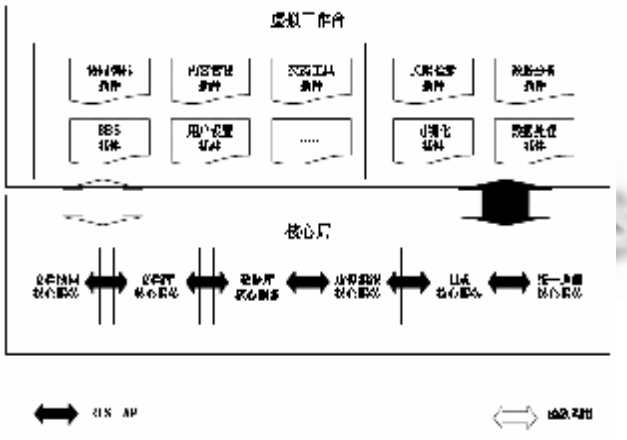

图 1　基于插件的 duckling 应用集成框架

2.1.3 DLM 框架介绍

DLM 使用 Portlet 技术[2]，是基于 Duckling 应用集成框架开发的数据集成和共享工具，是 BS 架构的 Web 应用[3]，提供通用的数据集成和共享功能。DLM 主要包括数据采集，数据处理，数据共享三个模块，如图 2 所示。

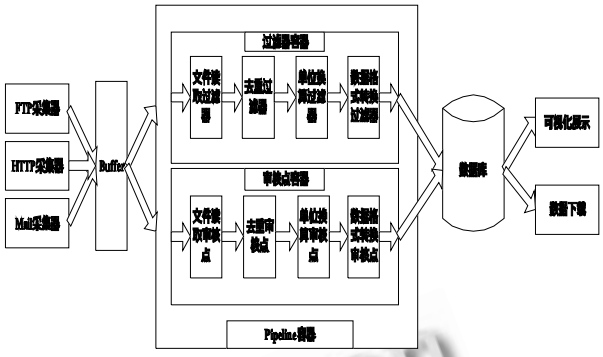

图 2　DLM 架构图

数据采集模块以 FTP、Email 等方式从多个站点采集数据，并将结果保存在缓存中。数据处理模块对数据进行处理，分为若干的过滤器或审核点，每个过滤器或审核点只实现了一个基本的操作如去重、单位换算、格式转换等，每个操作都是数据处理的一个必要步骤，最后将处理结果保存到数据库中。数据共享模块包括可视化展示模块和数据下载模块两个子模块。

DLM 数据集成和共享的流程是：数据采集模块采用多种方式将各个站点的原始数据采集到缓存中，触发数据处理模块。数据处理模块采用管道式处理流程，分为自动处理、手动处理、自动和手动处理结合三种方式。如果选择自动方式，用户通过 XML 文件对过滤器进行配置，配置好后就可以按照线性的顺序对每个过滤器自动触发执行；如果选择手动方式，用户通过 XML 文件对每个审核点进行配置，并以线性的顺序对数据加工过程中的每个步骤进行审核；如果用户选择半自动方式，则是上面两种方式的结合。数据共享模块提供数据的可视化页面展示和和数据下载两种共享方式，每种方式又可以提供多种方式的共享服务。

2.1.4 DLM 发布

启动 duckling，并将 DLM 部署到 tomcat 中。如图 3，点击最上面的“+”按钮新建一个页面，在右侧“可供加载的应用插件”中选择需要的插件应用，点击箭头朝左的按钮后添加，点击下面的“提交”按钮即可进行发布，点击下面的包含该插件的页面的链接即可访问。

2.1.5 小结

协同工作环境套件 Duckling 通过 Portlet 门户机制可以方便的将各类应用整合到统一框架之下，实现应用和资源的集成。DLM 将数据采集、处理和共享整

合为一个完整的流程，正是基于 Duckling 的这个优点。

图 3 Portlet 插件在 Duckling 中的发布

2.2 DLM 模块

DLM 提供通用的数据集成和共享功能：(1)数据采集：把分散的数据集中；(2)数据处理：把原始数据转换为规范化数据；(3)数据可视化：提供初步的数据展示界面；(4)数据下载：提供多种方式的数据下载功能。

2.2.1 数据采集模块

在数据采集模块中，数据采集方式有 FTP 和 Email 方式，每种方式都有一个监控器，以及一个 XML 配置文件。

(1) FTP 采集器

FTP 采集器通过 ftp 监控器(ftpmonitor)，来监控相关目录下的文件，并将文件上传到页面中。用户可以对参数进行配置，如选择要监控的目录，设置目录的扫描时间间隔等。可以启动多个 ftpmonitor 线程来监控多个站点的上传数据。此种方式可以很好的支持大数据量的上传和断点续传。

(2) Email 采集器

Email 采集器通过 email 监控器(emailmonitor)，来监控相关的邮箱。用户可以对邮箱信息进行配置，如用户名，密码、收邮件协议等。新邮件到达时，读取附件并上传到页面中。可以运行多个 emailmonitor，监控多个站点的上传数据。当配置了一个新的站点时，实时启动该站点的监控线程。

2.2.2 数据处理模块

DLM 的数据处理基于工作流形式[4]，其中的每个步骤是可以定制的。数据处理分为若干的过滤器或审核点[5]，每个过滤器或审核点只实现了一个基本的操作如去重、单位换算等。用户可以选择自动、手动等处理方式进行处理。

(1) 文件读取过滤器

要处理的文件类型，默认是所有格式。用户可以选择要进行过滤的文件的格式，即要处理哪种或哪些格式的文件，将得到的结果保存在缓存中，用来进行下一步的处理。

(2) 去重过滤器

上传的数据可能相同，这样就需要过滤出相同的数据。用户设置所需要的唯一键值，进行去重过滤。

(3) 单位换算过滤器

不同站点的数据可能使用不同的单位，这时就需要进行单位的转换。允许用户上载和选择自定义单位换算器。

(4) 数据格式转换过滤器

需要入库的某个字段，在不同站点的文件中的数据格式的表示可能不同，这时就需要进行数据格式的转换，方便后面的入库处理。

(5) 数据入库预处理过滤器

对访问数据库的特定代码进行封装，使数据处理的业务逻辑的细节问题与数据库分开，使代码能够最大化的得到重用[6]。可以使用 XML 文件或 properties 文件配置数据库连接时需要的信息。

2.2.3 数据展示模块

在数据的展示模块中，可以按照两种方式进行展示。

(1) 按课题展示数据

可以按照时间序列展示数据，使用 ChartDirector 来进行展示，基本的展示方式有折线图、饼图、柱状图、区域分布图、向量图、数据比较等方式。也可以使用 google 地图 api 的形式显示站点的状况和最新数据。

(2) 综合数据展示

对数据进行统计时，时间上可以按分、小时、天进行统计，实现多源数据的统一化或融合。虽然不同站点的数据是隔离入库的，但可以对同类数据进行对比或融合展示，可以用下拉列表的形式进行展示，也可以使用 google 地图进行展示。另外，DLM 还支持不同指标的对比。

2.2.4 数据下载模块

在数据的下载模块中，DLM 提供站点的规范化数据下载和整合打包方式，具体如下。

(1) 通过 FTP、HTTP、Email 的方式进行数据的共享，提供原始数据的下载。

(2) 通过选择站点、时间等下载数据，可以下载单站点、多站点的数据，也可以下载统计后的数据。

(3) 通过选择 google 地图上的站点下载数据，可以打包进行下载。

3 DLM关键技术

3.1 处理流程的可配置性

通过向导式配置，用户可以定义数据采集方式、去重过滤器、单位转换过滤器、是否进行数据审核等，并保存在 XML 配置文件中。每个站点保存成一个 XML 文件。下面是一个 XML 配置文件的简单示例，如图 4。

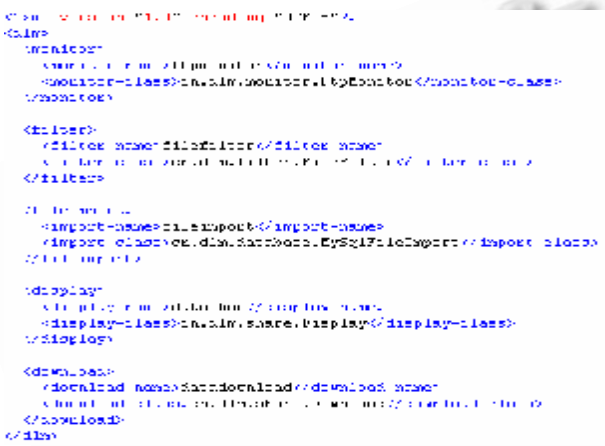

图 4　XML 配置文件示例

在该示例中，定义了一个采集方式为 ftp 方式的文件采集器，一个对文件类型进行过滤的文件读取过滤器，一个选择使用 MySQL 数据库的数据入库器，一个进行页面展示的数据展示器，以及一个提供数据下载的数据下载器。用户也可以配置更多类型的过滤器，满足实际的过滤需求。XML 常用的解析技术有 DOM(Document Object Model)，SAX(Simple API for XML)，DOM4J 和 JDOM 等几种[7]，每种技术都有其优点和缺点，适用于不同的应用，可以根据实际情况选择某种技术对 XML 文件进行解析[8]。当把 XML 文件配置好后，就可以进行数据的采集、处理、展示和下载了。

3.2 数据展示的实现

数据的展示可以用 ChartDirector 来展示，它采用多线程结构，拥有基于 API 的对象，允许用户控制和定制图表细节，提供直线图、曲线图、饼图、泡沫图、瀑布图、甘特图、向量图、数据对比等方式，特别适合应用于具有高性能要求的服务器端应用程序开发。

数据的展示也可以使用 google map api 来实现，它支持自定义扩展地图类型，有丰富的图形标注工具，有多种数据接口。使用它可以展示某个站点的最新数据，也可以展示多个站点的运行状态，如正常、故障等。可以按照课题来展示数据，也可以对数据进行综合统计后进行展示。

4 DLM的应用

目前，该工具已经在科研活动中得到了实际应用。在大气科学领域，需要从多个观测站点进行大气观测数据的采集，把数据处理后保存到数据库中，进行页面的可视化展示，并提供数据下载。在分子生物学领域，需要采集基因序列、蛋白质肽段序列等数据，对数据进行去重等处理，将结果保存到数据库中，最后以页面的形式进行展示。如下图 5 所示，可以按照图中的步骤新建一个数据处理流程。

图 5　DLM 在科研活动中的应用

首先，在输入流程信息这个选项中，需要输入流程的名字和流程描述。下一步，选择采集方式，可以选择 FTP 和 Email 两种采集方式，每个站点可以选择一种采集方式。下一步，设置处理单元，选择是否需要进行人工审核，如果选择该项，新数据需要经过审核后才能入库。下一步，设置使用方式，提供数据下载和页面数据展示两种方式，需要选择至少一种方式。这样一个站点的处理流程就建好了。点击启动按钮，这样就可以对该站点进行数据采集、处理和展示了。

(下转第 161 页)

(上接第 252 页)

5 总结

本文针对虚拟科研应用环境中遇到的数据集成和共享问题，实现了基于虚拟科研环境协同工作套件 Duckling 的数据集成与共享工具 DLM，可以通过 XML 文件对数据处理流程进行配置，配置好后可以实现站点数据的采集、处理和展示功能，具有很好的通用性和定制性。该工具使协同工作套件 duckling 能更好的支持科研活动，并在科研活动中得到了实际应用。

参考文献

1 中国科学院计算机网络信息中心协同工作环境研究中心.为什么选择协同工作环境套件(Duckling). 2010-02-20,http://duckling.escience.cn/.

2 马增辉,解建仓,周永进,罗军刚.一种多 Portlet 之间交互方法的研究.计算机工程, 2007,33(13):242—244.

3 朱韶平.XML Web Services 技术在工作流访问代理协同接口中的应用.科学技术与工程, 2008, 8(3):805—807.

4 李军怀,张彤,张景,刘海玲.一种基于 XML 的工作流过程定义语言研究与应用.计算机工程, 2005, 31(15):53—55.

5 李建.Java Web 开发中过滤器组件应用及实例解析.电脑开发与应用, 2009, 22(11):58—60.

6 霍志华,王建林,薛尧予.基于工作流和 XML 的生产报表系统设计与实现.计算机工程与设计, 2008, 29(16):4249—4251.

7 陈小毛,汤文兵. Java 解析 XML 的方法比较研究.中国新技术新产品, 2009,(15):25—25.

8 刘芳,肖铁军.XML 应用的基石：XML 解析技术.计算机工程与设计, 2005,26(10):2823—2824.