农业数据网格体系研究

吴华瑞，孙　想
（国家农业信息化工程技术研究中心，北京 100097）

摘　要：由于农业数据具有地理分布、异构和动态等特性，网络环境下农业数据的分发组织和高效能处理是一个非常复杂且具有挑战性的问题。该文设计了层次式的农业数据网格系统软件体系结构；研究了基于时空本体的农业元数据建模、农业数据空间聚类与映射、多系统信息协同和层次访问控制机制等关键技术；构建了农业数据网格节点应用环境。测试检验结果表明，所设计的农业数据网格体系具有良好的网络高性能传输能力。

关键词：数据网格；网格体系；农业数据；元数据；信息协同
中图分类号：TP393；TP274.2　　**文献标识码**：A　　**文章编号**：1002-6819(2006)11-01830-04

吴华瑞，孙　想. 农业数据网格体系研究[J]. 农业工程学报，2006，22(11)：183- 186.
Wu Huarui，Sun Xiang. Architecture of agricultural data grid system[J]. Transactions of the CSAE，2006，22(11)：183-186.（in Chinese with English abstract）

0　引　言

由于农业技术和计算机技术的发展，现代大型农业科学工程研究、农业信息服务和数字媒体应用中的数据呈爆炸式增长，农业数据已经成为一个重要的资源，例如：全球气候模拟、精准农业、遥感数据、生物计算、作物生长模拟、土壤数据、水利资源、电子农务、电子政务、数字媒体等应用，它们的数据量将达到几十个 TeraByte 至 PetaByte 的级别，地理上广泛分布的用户都希望能够访问、分析和使用这些庞大的分布数据，而他们的分析方法大都计算复杂且计算量大，这种结合海量数据集合、地理上分布的用户和资源，以及计算密集型的分析处理应用导致了现有的数据管理体系结构、方法和技术已经不能满足高性能、大容量分布存储和分布处理能力的要求，如何存储、分发、组织和管理、高性能处理、分析和挖掘海量分布数据成为许多应用的首要问题[1]。网格技术的发展为解决这个问题提供了有效的技术途径，本文针对建立国家级农业数据网格系统所涉及的理论和应用基础问题，提出了一种层次式的农业数据网格系统软件体系结构设计方案，并深入研究了农业数据网格涉及的主要关键技术，初步建立了农业数据网格节点试验应用环境，来方便管理分布异构存储的海量农业数据资源，并有效进行各类资源的优化调度和远程应用。

1　系统体系结构与设计

农业数据网格体系结构设计为 3 层结构（图 1）。

第一层网格基础服务环境，包括资源聚合器、数据代理服务、安全服务和系统管理等。资源聚合器主要功能是对计算和设备资源的接入、监控和调度进行管理，

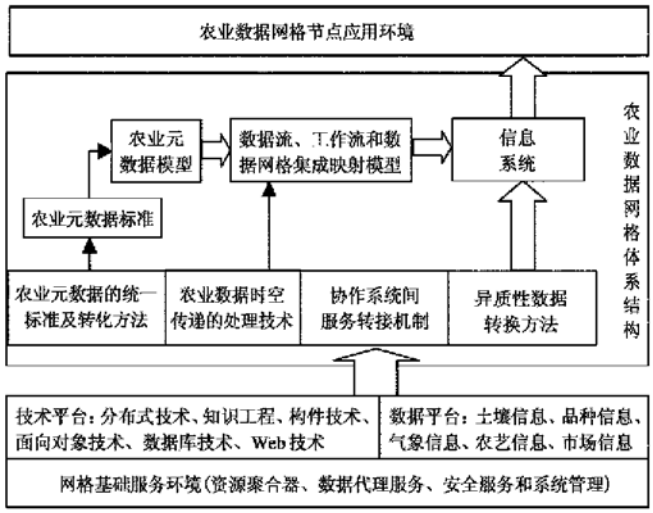

图 1　农业数据网格体系结构

Fig. 1　Architecture of agricultural data grid system

支持高性能计算所需功能；数据代理服务模块主要提供数据的访问优化、调度和服务，管理分布异构存储资源上的数据为一体，提供数据的统一访问，允许数据的高速传送、复制操作和副本管理，以及虚拟数据管理；安全服务主要支持单一登录认证和多层次的访问控制和授权机制[2]。系统管理主要实现网格系统用户的建立和删除，系统的配置和部署，以及全网格系统状态的监控。农业数据网格基础服务环境的构建遵循 OGSA 的标准，借鉴成熟的 Globus Toolkits 2.0 的体系结构，这里不再赘述。

第二层是农业数据网格业务核心层，主要包括农业元数据的表示与转化、数据时空传递处理、系统协作服务转接和异质性数据转换方法等。农业数据网格业务核心层是一个一体化网格数据管理与服务容器，它定义了一套农业数据管理协议，并提供相应的农业数据传输、管理、控制和访问功能，以数据管理为中心，面向上层应用提供通用、可靠的服务，面向底层则将网格中各种资源紧密结合起来，为用户提供一体化高性能计算服务和

收稿日期：2006-08-19　修订日期：2006-11-08
基金项目：国家科技攻关计划项目(2005BA113A03)；农业科技成果转化资金项目(04EFN211100002)
作者简介：吴华瑞(1975-)，男，山东冠县人，博士生，主要研究方向为数据网格、中间件技术。北京　北京农业信息技术研究中心，100097。Email：wuhr@nercita.org.cn

信息处理服务[3]。

第三层是农业数据网格节点的应用环境,它基于网格中结点主机提供的核心业务功能,面向整个网格提供作业管理与服务[4]。

农业数据网格的各个模块既独立又有联系,模块之间通过协同服务机制整合在一起,只要是符合标准的数据和系统都可以无缝的整合到整个网格中,从而实现功能一体化、权限管理同步化和 SSO 单点登录。

2 农业数据网格实现的关键技术

2.1 基于时空本体的农业元数据建模

农业数据的一个显著特点是时间性和空间性,同样的农业数据在不同时间不同地域的应用千差万别,因此,农业数据需要考虑时空性。基于时空本体的农业元数据建模就是采用 WEB 本体语言和语义网规则语言(SWRL, Semantic Web Rule Language)描述时空动态分布的农业数据的时间原语和空间原语[5],根据农业数据资源的内涵(如规则集、多媒体集、语义片断集、线性规划模型等)确定数据资源的层次结构;采用多维结构描述数据本体及其属性(如数据的基本内容、数据类型、数据表示形式、数据所属领域等)。基于时空本体的农业元数据模型(图 2)建立过程如下:

• 农业数据资源内涵的描述与分析,确定应用的范围,通过字典定义需要管理的数据;

• 确定农业数据的层次结构,自顶向下进行划分为概念级、公理级、规则级和方法级;

• 设计农业数据空间结构,确定其内容、类型、表示形式、数据所属领域和位置坐标;

• 依据时空本体单元连接具有公共坐标轴的数据资源空间,形成统一的高维农业元数据模型,通过它来判别农业数据网格体系中数据来源以及相应关系的判别;

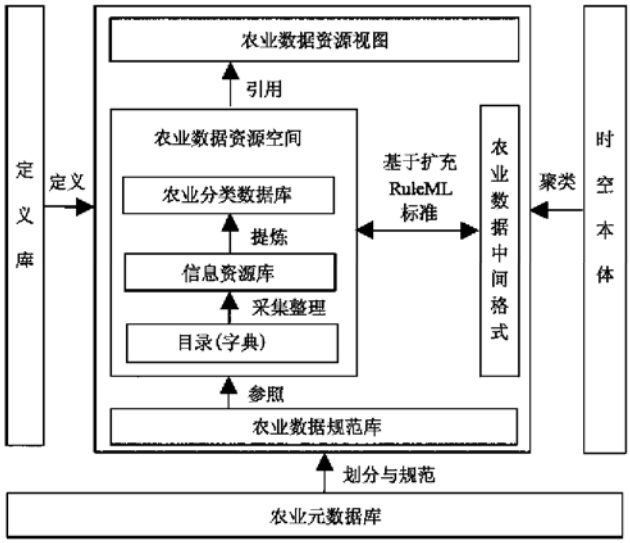

图 2 农业元数据模型

Fig. 2 Agricultural meta-data model

• 确定基于扩充 RuleML 标准 BNF 的数据中间格式,解决 RuleML 规范不能表达模糊数据和模型计算,前提项中不支持 OR 关系运算的问题,解决异构数据在网格环境下的协同。

2.2 农业数据区域空间聚类与映射模型

在工作流中的定义阶段确定用户的数据需求,包括名称、类别、描述、需求人员、所属活动及相应的时间约束等。时间约束与活动执行中的时间约束相关联,以实现与活动执行相同步的数据服务。数据需求与角色之间是多对多的关系,即一个角色在同一活动中可以具有多个数据需求,一个数据需求可以为多个角色所拥有[6]。数据需求可以在工作流的定义阶段基于组织经验和最佳实践而指定,也可以在工作流的执行阶段由具体的参与人员提出。在数据源和数据接受者之间建立关联,建立区域时空数据的聚类和分类模型,实现容易理解的高维数据资源聚类和分类,方便数据流的快速正确传递。通过相应的控制机制对相关数据流结构和有关属性进行调整,保证数据流的有效性和对环境变化的自适应性,最终形成农业地理分布的数据流、工作流和数据网格集成映射模型[7]。即第一层映射把数据源映射到数据流的节点,第 2 层映射把每个角色映射到工作流的活动,第 3 层映射把每个角色映射到网格上接收成员[8],见图 3。

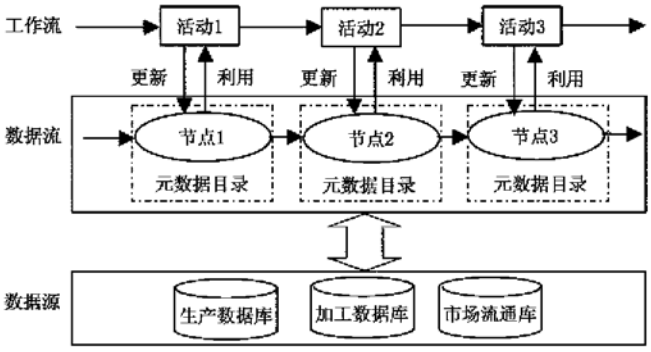

图 3 数据流与工作流集成模型

Fig. 3 Integrated model of dataflow and workflow

2.3 农业数据的统一访问接口

农业数据网格为网格用户提供一套统一的访问 API 函数和软件包,这套 API 将农业数据网格内部数据结构的复杂性和操作的复杂性掩藏起来。数据使用者采用类似于 NFS 文件系统的访问接口对各种异构存储系统进行访问,系统则根据数据存储特性采用相应的数据访问协议代理用户对底层存储资源数据进行操作,通过这套接口可以对数据进行访问,进行各种管理[4]。也可以进行应用开发。准备设计其访问过程为:

• 数据使用者使用数据访问 API 向数据代理服务 DRB(data Request Broker)提出数据请求,例如 Open, Read, Write, Move 等;

• DRB 监听请求后,派生 Proxy 线程,对该请求进行解析和处理;

• Proxy 查询元信息服务器, 获得数据的存储信息和相应访问方式的元信息;

• Proxy 根据返回信息调用对应的访问接口获取数据, 将结果返回给数据使用者。

多个 DRB 之间组合成为联邦 DRB, 形成一个功能更加强大的数据服务器来响应外界的数据服务请求。用户请求可以发送到任何一个 DRB, 只要给出某个特定数据集的标识, 农业数据服务器就会从该 DRB 调度到某个合适的 DRB 上, 在该 DRB 上生成 Proxy 提供数据访问操作服务, 并能够协调其它 DRB 联合提供服务。这种机制比单 DRB 服务可靠性更好、功能更强, 更容易实现系统的扩展。

2.4　多个农业信息系统的协同机制

通过建立系统服务转接机制, 保证核心节点与各子节点, 各子节点之间的实时通信联系, 各个服务系统相互之间也可以交流传递数据。在用户登录后, 可以访问数据网格中的所有系统和资源, 整个共享服务网络都可以为这个用户提供服务, 用户可以在不同的服务系统间自由穿行。本文研究的农业信息系统服务转接分以下 3 种(图 4) :

1) 在一个子节点上对于用户的请求查询不到答案时, 决策模块根据信息列表, 按照相应转接协议将该请求转接到其他节点上进行查询, 找到答案后再返回给用户。在每个协作节点上都有整个数据网格的资源元数据目录结构及其元数据列表。

2) 当用户在一个服务平台上访问的数量超过了额定数量后, 将进行用户服务的转移。由决策模块对新的用户请求进行筛选。

3) 当用户登录的节点由于网络阻塞访问速度缓慢, 这时将进行用户服务的转移。由决策模块根据各个子节点和用户数据, 以及网络状况, 自动挑选一个距离用户的网络最通畅的节点, 把用户的请求、以及用户数据按照转接协议都打包转移过去, 由新的节点继续为用户提供服务。

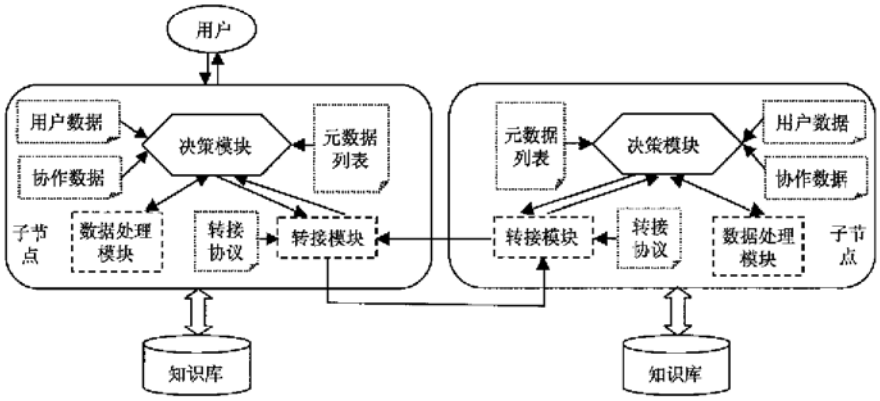

图 4　农业数据网格各节点间的服务转接图

Fig. 4　Service-transfer among nodes of agricultural data grid

2.5　层次访问控制机制

农业数据网格系统准备采用基于角色的分层次访问控制机制(Role-Based Access Control, RBAC) 。所谓角色是指与特定操作活动相关的一组动作和权限集合。系统管理员只需根据应用特点, 依据某些授权原则建立相应的角色。对于具体的网格用户, 系统管理员根据需要对其授予某个或者某些角色, 使其获得该角色所定义的操作。这种机制自主性适中, 且便于权限的发放与回收。访问控制分为两层。在网格全局用户映射到局部用户之前, 系统需要对全局用户的授权进行验证, 只有拥有合法权限的用户才能映射到数据资源所在局部系统。局部系统还必须使用它自身的局部访问控制机制对授权再次进行验证。只有完全通过这两步验证才能对数据进行访问操作[9] 。

3　农业数据网格节点的应用试验环境

基于以上研究的农业数据网格体系结构, 搭建农业数据网格节点应用试验环境。各个数据网格节点上的农业信息资源、农业决策支持系统通过农业数据 DRB 服务容器进行统一格式转换, 通过数据管理中间件对以上资源和服务进行分类和映射; 基于农业数据网格核心业务层实现任务的调度、数据管理和可信计算功能, 最终将信息服务按需分配到各个网格节点上的用户。农业数据网格节点部署架构如图 5 所示。

为了评估网格环境运行效果, 设计了一个模拟程序, 模拟的网格环境由 9 组服务器主机和 5 个数据源组成。模拟应用由番茄、黄瓜和奶牛等 3 个农业智能决策系统的 50 个任务组成, 每个任务至少存取一个数据源来获得输入数据, 任务的运行时间确定。对农业数据网格节点应用试验环境下的数据传输性能与传统 B/S 模式下的分布数据传输性能进行对比测试, 测试结果如图 6 所示, 在同等服务器群组环境下, 网格环境下的数据传输性能可达到 184. 93M Bps, 而传统 B/S 环境下的数据传输性能仅仅达到 161. 63M Bps, 这对于大规模农业数据的网络高性能传输无疑是个很大的突破。

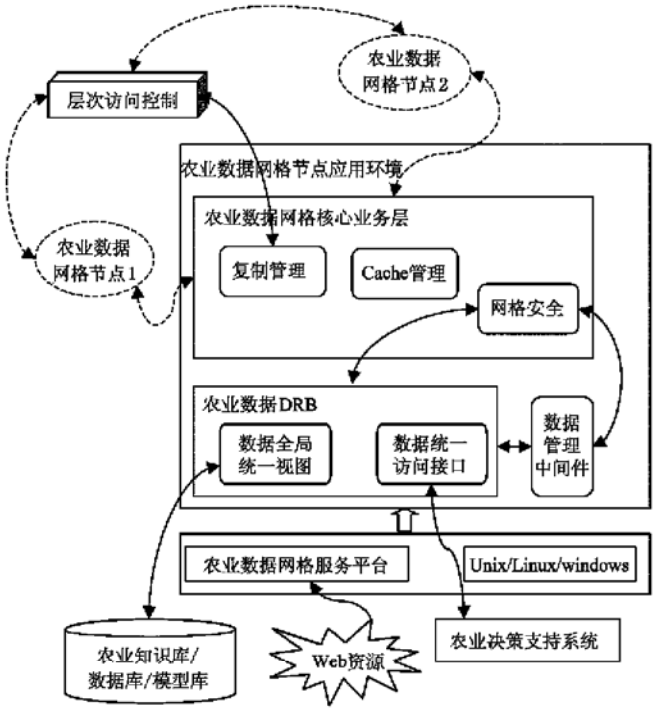

图 5 国家农业数据网格应用节点部署架构

Fig. 5 Architecture of application nodes for national agricultural data grid

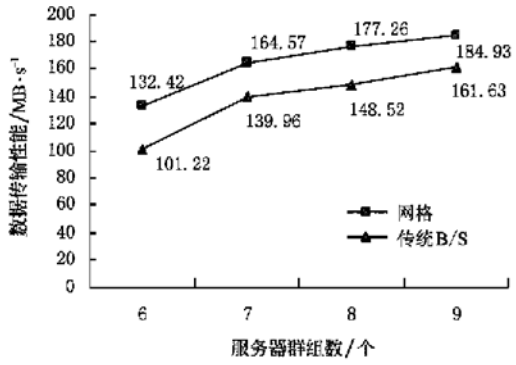

图 6 农业数据网格传输性能测试结果

Fig. 6 Results of performance test in agricultural data grid

4 结 论

目前网格技术的发展变化很快,国际上还没有成型的网格标准,本研究设计实现了具有分布协同计算和交叉并行服务能力的农业数据网格体系,为农业数据资源的共享,农业信息的互通互联,以及农业信息系统协同工作提供了新型方法和技术支撑,实现了网格技术在农业领域应用的突破,从而促进网格基础服务软件向行业应用领域发展。但是,如何建立农业网格环境下多源农业信息可信度评价模型和系统安全监控和跟踪机制,以保证农业数据网格环境下数据传输和决策业务各个环节中的机密性、可鉴别性、可控性和可用性不受破坏,尚需进一步研究。

[参 考 文 献]

[1] 洪学海,许卓群,丁文魁. 网格计算技术及应用综述[J]. 计算机科学, 2003, 30(8): 1- 5.

[2] William H Bell1, Diana Bosio. Project spitfire-towards grid webService databases[J]. Global Grid Forum 5, Edinburgh, Scotland, 2002: 21- 24.

[3] Zhuge H. China's E-science knowledge grid environment [J]. IEEE Intelligent Systems, 2004, 19: 13- 17.

[4] Chervenak A, Foster I, Kesselman C, et al. The data grid. Towards an architecture for the distributed management and analysis of large scientific datasets[J]. Journal of Network and Computer Applications, 1999: 126- 131.

[5] 刘大有,胡 鹏,王生生,等. 时空推理研究进展[J]. 软件学报, 2004, 15(8): 1141- 1149.

[6] Bester J, Foster I, Kesselman C, et al. GASS. A data movement and access service for wide area computing systems[M]. ACM Press, Atlanta, GA. 1999: 78- 88.

[7] Zhuge H. Resource space model, its design method and applications[J]. Journal of Systems and Software, 2004, 72(1): 71- 81.

[8] 丁 箐,陈国良,顾 钧. 计算网格环境下一个统一的资源映射策略[J]. 软件学报, 2002, 12(07): 1303- 1308.

[9] 肖 侬,任 浩,徐志伟,等. 基于资源目录技术的网格系统软件设计与实现[J] . 计算机研究与发展, 2002, 39(8): 902- 906.

Architecture of agricultural data grid system

Wu Huarui, Sun Xiang

(*National Engineering Research Center for Information Technology in Agriculture, Beijing* 100097, *China*)

Abstract: High performance processing of agricultural data in the grid environment is a complex and difficult problem, mainly due to agricultural data geographic distribution, heterogeneity and variety. A multi-layered software architecture for agricultural data grid system was designed. Agricultural meta-data model based spatio-temporal ontology, spatial clustering and mapping for agricultural data, information cooperation, role-based access control were researched. The application environment of agricultural data grid node was constructed. The inspection experiment indicates that the agricultural data grid system provides high performance computation and communication capability.

Key words: data grid; grid system; agricultural data; metadata; information cooperation