

基于云计算的农业大数据共享服务平台研发*

余万民¹, 范蓓蕾^{2*}, 钱建平³

(1. 贵州省农业信息中心, 贵阳 555001; 2. 中国农业科学院农业信息研究所, 北京 100081;

3. 中国农业科学院农业资源与农业区划研究所, 北京 100081)

摘要:【目的】为了整合多源、异构、海量农业数据, 基于云计算等技术设计并实现农业大数据开放共享平台, 有利于农业数据资源的充分利用。【方法】文章依据科学合理标准规范, 以农业大数据资源编目元数据、大数据整合集成与建库、农业大数据共享服务为主线, 利用 Hadoop 大数据技术, 基于 J2EE 平台和构件化开发思路, 构建农业大数据共享服务平台。【结果】该平台包括农业大数据集成管理、农业大数据共享服务管理和元数据管理, 其中农业大数据集成管理实现了对各渠道所得农业数据进行整合集成、数据质量控制和更新管理, 农业大数据共享服务子系统实现了对分散、多源、异构农业非空间数据的目录导航、查询检索与共享交换等服务, 是用户查询访问和使用农业大数据的统一入口, 元数据管理子系统实现了对多源大数据的统一描述、发布管理与权限控制。【结论】该平台在贵州省进行了开发应用, 实现了贵州省农业数据资源的统一集成、统一管理、统一共享和统一服务, 为贵州省农业云服务及其各类应用提供了数据基础。

关键词: 农业大数据; 资源目录; 云计算平台; 数据共享

DOI: 10.12105/j.issn.1672-0423.20200103

0 引言

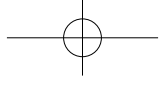
大数据是具有海量、快速、多样、真实、高价值等特征的信息资产, 需要先进的处理技术进行采集、存储、分发、管理和分析, 从而提供更强决策力和洞察力^[1]。大数据现象源于互联网及信息技术的广泛应用, 大数据概念体系在 2014 年左右逐渐成形, 发展出相关的技术、产品、应用和标准, 并逐步形成包括数据资源与 API、开源平台与工具、数据基础设施、数据分析、数据应用等内容的大数据生态系统^[2]。大数据技术已成功地应用于银行、保险、在线用户行为理解和个性化以及环境研究等行业, 但在农业领域于近几年才开始, 并未得到广泛应用^[3]。农业大数据为现代农业发展带来巨大机遇。我国农业农村信息化工作经过近十多年的发展, 信息化基础设施已有较大改善, 广播、电视、电话、互联网等纷纷入村入户, 智能手机、农业物联网、智能农机具逐渐发展并普及, 为农业大数据的发展应用提供了坚实的基础^[4]。以大数据为代表的现代信息技术

收稿日期: 2020-01-20

第一作者简介: 余万民 (1991—), 高级工程师。研究方向: 农业信息技术。Email: 878456287@qq.com

*通信作者简介: 范蓓蕾 (1981—), 博士、高级工程师。研究方向: 农业信息技术。Email: fanfanbl@163.com

*基金项目: 中国农业科学院科技创新工程引进英才“智慧农业”(962-3)



推动了农业生产方式、流通方式、贸易方式的变革^[5-6]。

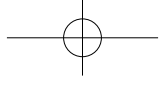
2020 年“中央一号文件”提出依托现有资源建设农业农村大数据中心，加快物联网、大数据、区块链、人工智能、第五代移动通信网络、智慧气象等现代信息技术在农业领域的应用。近年来，我国在部分地区成功开展了多个农业大数据平台试点，如全国基层农技推广信息化平台^[7]、国家农业科技云平台、渤海粮仓科技示范工程大数据平台^[8]、农业物联网大数据共享及综合应用平台^[9]，为农业生产和研究提供了不同层面的农业数据支撑和服务。农业物联网、农业信息系统、农业云服务、农业卫星遥感等软硬件平台在农业领域的不断投入和使用，促进了我国农业数据的爆炸式增长^[10]。农业大数据来源广泛、更新频繁、内容丰富、结构复杂，从内容上看，既包括农业生态资源、农业生态环境、农业生物资源等数据，又包括农资及农产品的生产、加工、运输、储藏、流通、销售、质量安全等数据；从结构上看，包括结构化数据，非结构化数据以及半结构化数据；从格式上看，既有文本、图像、视频、音频等数据，也有文档、文件、数据库等资源数据^[11]。

对农业领域来说，组织、分析和解释各种农业数据非常重要。国内外专家在农业大数据资源管理和应用上进行了大量研究。Perakis 等设计了一种应用于精准农业和精准畜牧业的无缝访问大量不同类型、不同来源超大规模数据集的平台，提出一种对大数据发现、处理、组合和可视化服务的大规模高性能计算基础设施的安全和非中介访问方案，支持对传感器数据、环境和气候历史数据、卫星和航空图像等数据的集成和访问，不仅能为农民提供服务，也能为农业生产链条上的其他主体如企业、技术人员等提供服务^[12]。Shrivastava 和 Marshall-Colon 提出通过分析各种来源的农业大数据，能够加速育种计划，有助于设计适应复杂环境的高产作物类型，从而解决粮食安全问题^[13]。Sabarina 和 Priya 提出一种基于张量的特征约简模型，用于对农业大数据分析时降低数据维度，从而减少数据分析时间^[14]。Huang 等分析整理了现有的遥感数据资源和管理发展，并提出了一种 FLTL (Four Layer Twelve Level) 遥感数据管理结构，用于管理和应用农业遥感大数据^[15]。李艳和刘成龙基于 Hadoop 框架设计了农业大数据挖掘系统，通过分布式数据组织方式实现了对农业大数据的高效率分析、计算和处理^[16]。钱晔等基于农业大数据中心积累的数据成果，通过人工神经网络算法构建鲜切花价格预测和质量等级识别模型，实现了对农业大数据的智能化应用^[17]。

通过对已有研究分析可知，农业大数据应用已取得一定成果，但农业大数据资源分散，不同平台、部门之间共享困难，而农业大数据资源集成、共享、交换等方面的研究较少。文章针对贵州省农业大数据资源共享管理问题，具体分析农业大数据的来源和内容，从多源数据集成、元数据管理和数据共享交换 3 个方面开展农业大数据共享服务平台的研究与设计，为农业大数据的开放共享提供一套切实可行的技术方案，促进农业数据资源的高效利用。

1 需求分析

深入调研贵州省农业数据资源基础以及业务需求情况，从数据基础和功能需求两方面梳理分析大数据资源管理平台的需求。



2020 年 2 月

1.1 数据基础

农业各类数据通常来自各类农业政务服务系统、农业产业服务系统、物联网、互联网等。从业务种类区分, 农业数据可以分为农业基础条件数据(农业区划、农业气象、土壤、耕地、草地、水域等)、农业资源要素数据(种植、经济作物、畜牧、渔业水产、高标准农田、农业园区等)、农业经营交易数据(农资、新型经营主体、农产品贸易、农产品价格、农田用途、农村土地确权、流转等)和农业管理服务数据(水肥管理、病虫害防治、粮食补贴、三品一标、农产品溯源、农业专家库、农民培训等)等。从数据类型来看, 农业数据包括关系型数据库、文档、图片、流媒体等多种格式。各类数据资源分散在不同的部门, 部署在不同的服务器或云平台上, 对数据进行采集汇聚有多种渠道, 主要的方法有以下 3 种。

(1) 数据共享接口: 主要适用于农业业务部门内部系统或可以提供数据共享服务的系统平台。

(2) 物联网数据汇聚: 主要适用于物联网设备产生的数据, 可通过物联网支撑平台汇聚后发送到大数据平台上统一汇聚。

(3) 爬虫采集系统: 主要适用于互联网数据的采集汇聚, 如互联网上农业相关的各种舆情信息、农产品价格等。

1.2 功能需求

农业大数据共享服务平台的目的是将各种农业系统和平台产生的农业数据进行汇总和集成管理, 并向需要数据的用户提供数据共享和交换服务, 其面向的用户有公众用户、政府用户和企业用户 3 种。根据平台目标, 农业大数据共享服务平台可以划分为 3 个主要功能, 分别是农业大数据集成管理、农业大数据共享服务和元数据管理。

(1) 农业大数据集成管理通过各种技术手段, 将不同平台和数据库的农业数据资源进行汇聚, 并提供对数据清洗、转换、质量管理等服务。

(2) 农业大数据共享服务提供数据共享和交换功能, 根据用户不同, 应该提供不同级别的数据共享。

(3) 元数据管理则是将集成汇聚的数据进行资源编目, 从海量数据中抽取描述数据内容的元数据, 使数据管理更加规范和容易。

2 农业大数据共享服务平台架构

为了与其它已有系统实现无缝衔接, 该平台采用基于云计算的全网络化 B/S 系统架构。农业大数据共享服务平台架构如图 1 所示。该平台采用 Hadoop 框架进行设计, 最底层是数据存储服务, 为了满足多种结构类型, 采用关系型数据库如 Oracle、Mysql、Sql server 等管理结构化数据, 采用 Hbase 和 Redis 管理非结构化数据, 采用 HDFS 管理分布式文件系统, 采用 FastDFS 管理磁盘文件系统, 采用 Lucene 管理异构索引库; 之后, 进入数据共享交换层, 通过 ETL、WebService、FTP 等技术, 将各业务部门的数据进行汇总, 编制资源目录, 建立元数据库; 在此基础上, 对汇总数据云计算服务, 包括数据

仓库的构建、数据挖掘等服务；在此基础上提供应用服务，进行任务调度、数据检索、缓存服务和分库分表等，并在此基础上开发大数据资源管理平台，将汇总后的农业大数据根据不同级别、不同主体进行开放共享，为各级部门提供更加便利的数据服务。

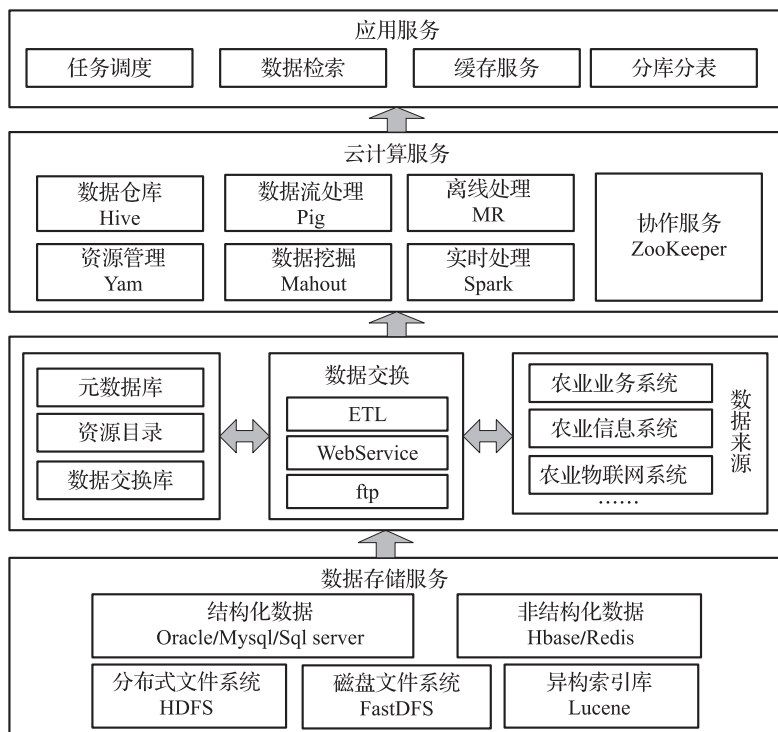


图 1 农业大数据共享服务平台框架

Fig.1 Framework of agricultural big data sharing service platform

3 功能设计

根据功能需求分析，将农业大数据管理平台划分为 3 个子系统，分别是农业大数据集成管理子系统、农业大数据共享服务子系统和元数据子系统，如图 2 所示。

3.1 农业大数据集成管理子系统

该子系统实现从各渠道得到的农业基础、产业、监管、政务数据的整合集成、数据质量控制、更新管理等，主要提供数据整合、数据质控、数据封装、数据统计、集成规则管理等功能。

（1）数据整合：通过数据分类、数据同步、数据抽取、数据清洗等处理，实现不同来源、类型数据的整合集成。将敏感数据和普通数据进行区分，对数据按照不同安全级别进行分类管理。

（2）数据质控：对汇聚的数据提供质量监控管理，监控汇聚过程、交换接口、服务状态，定义“正则表达式”和数据字典，对数据质量进行校核，对集成的数据资源进行质量检验、异常和冲突处理。

2020 年 2 月

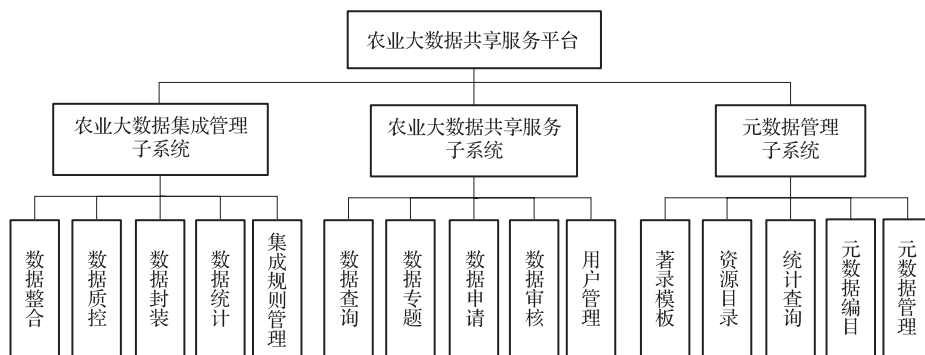


图 2 农业大数据共享服务平台功能结构

Fig.2 Function structure of agricultural big data sharing service platform

(3) 数据封装：对数据按照不同的格式、类型进行转换，按照不同的结构进行处理，将结构化数据存储到关系型数据库中，将非结构化数据存储到文件服务器上，并记录文件信息，确保数据的完整性。

(4) 数据统计：按部门或单位、数据分类、更新时间、农业要素属性等对集成的数据以图表等多种形式进行各类统计汇总和变化分析，包括单要素或多要素时间序列变化分析、区域构成分析、区域对比分析等，通过统计分析实现农业生产条件、资源要素、经营权属和管理服务情况。

(5) 集成规则管理：设置各类数据源集成、清洗、转换规则，包括数据同步、数据抽取、数据转换规则，以及数据质量检验和异常处理规则等；对平台所涉及到的用户访问权限代码、数据分类编码、数据属性值代码等数据字典和代码进行管理。

3.2 农业大数据共享服务子系统

该子系统实现分散、多源、异构农业非空间数据的目录导航、查询检索与共享交换等服务，主要包括数据查询、数据专题、数据申请、数据审核以及用户管理等功能，是用户查询访问和使用农业大数据的统一入口。

(1) 数据查询：提供基于关键词的简单查询、多条件组合的复杂查询、全文检索和关联图检索。对查询结果提供浏览功能，包括属性表浏览、空间数据浏览、底图展示以及元数据回溯等。

(2) 数据专题：提供基于不同主题的数据分析和展示服务，如农业生产、农民专业合作社、农产品市场价格等专题的分析图、分析表格等处理。

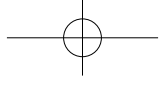
(3) 数据申请：针对用户提交的数据申请，提供审核与批复功能，如果审核通过，用户可以获得数据查看权限。

(4) 数据审核：对数据申请进行审核、开放共享端口等处理。

(5) 用户管理：提供用户注册、用户角色划分、用户权限设置等功能，实现对用户的统一管理。

3.3 元数据管理子系统

该子系统提供了对多源大数据的统一描述、发布管理与权限控制，主要包括著录模板管理、资源目录管理、统计查询、元数据编目以及元数据管理等。



(1) 著录模板管理：提供定制元数据模板的功能，也可以将指定元数据保存为元数据模板，对模板进行修改和删除等管理，各级用户通过元数据模板利用元数据模板可以实现元数据的快速录入。

(2) 资源目录管理：包括资源目录分类、资源名称、资源属性、数据提交形式、共享方式、更新途径、更新周期、资源数量级、责任部门等内容管理。

(3) 统计查询：按责任者、共享类型、数据分类、格式等条件进行元数据查询，对元数据进行统计汇总。

(4) 元数据编目：对元数据编目要素进行管理，按共享类型、开放方式、责任者等进行编目，按指定格式输出元数据目录，依据权限访问关联的数据。

(5) 元数据管理：依据管理权限，对元数据进行修改、删除、审核、发布等处理。

4 平台实现

农业大数据共享服务平台基于 J2EE 平台进行开发，利用 Spring boot、Duboo、Zookeeper 搭建的共享服务平台，用于 RESTful、XML/JSON 服务的快速开发、注册、发现、路由等工作；利用 Hadoop 大数据体系技术，对海量农业数据进行并行计算和分布式处理；支持结构化数据库系统 Oracle\MySQL\Sqlserver、利用 NoSql 数据库作为前端界面缓存数据库；同时构建文件库和异构索引库存储非结构化数据信息。

基于该平台架构和设计思路，在贵州省进行了开发应用，图 3 为该平台进行数据分类的界面，图 4 为平台子系统农业大数据集成管理中进行数据转换的界面。该平台将贵州省不同来源的农业资源数据进行了有效整合，并提供了统一的门户网站进行共享，提高了数据利用效率。



图 3 贵州省农业大数据共享服务平台应用界面

Fig.3 Application interface of Guizhou agricultural big data sharing service platform

2020年2月



图4 农业大数据集成管理子系统应用界面

Fig.4 Application interface of agricultural big data integrated management subsystem

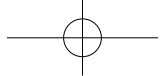
5 结论

针对农业数据格式多样、物理分散、不易利用等问题,采用大数据、云计算等技术,以元数据管理为基础,开发了面向多源异构农业资源数据整合应用的农业资源数据共享服务平台,打破了不同农业系统、数据库系统之间的壁垒,实现了数据的有效共享和利用。该平台在贵州省进行了开发应用,集成了农业部门内外多种数据源,包括各种农产品价格、农业生产、农村生产合作社、农业生产经济、农业脱贫工作等资源,通过提供专题服务,可以快速将多源数据进行集成,为农业主管部门提供各种资源的统计汇总“一张图”服务,实现对地区农业产业发展、农业经济等信息的快速全面掌握,为制定政策和发展计划提供了良好的数据支撑。

农业资源数据共享服务平台提供了数据集成、数据共享和元数据服务,平台功能以构件方式开发实现,有利于在该平台基础上进一步拓展开发业务功能。

参考文献

- [1] Gandomi A., Haider M..Beyond the hype: Big data concepts, methods, and analytics.*International Journal of Information Management*, 2015, 35(2): 137-144.
- [2] Viktor Mayer-Schönberger. 大数据时代. 杭州: 浙江人民出版社, 2012.
- [3] Kamilaris A., Kartakoullis A., Prenafeta-Boldú F.X..A review on the practice of big data analysis in agriculture.*Computers and Electronics in Agriculture*, 2017(143): 23-37.
- [4] 郝炘, 李建华, 牛明雷, 等. 物联网背景下农业信息化建设的现状、问题与对策研究. 农学学报, 2019, 9(11): 67-71.
- [5] 吴文斌, 史云, 段玉林, 等. 天空地遥感大数据赋能果园生产精准管理. 中国农业信息, 2019, 31(4): 1-9.
- [6] Mario L., Jorge E.H., Maria M.E.A.D., et al.Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture.*Computers in Industry*, 2020(117): 103187.
- [7] 卢文林, 李秀峰. 全国基层农技推广信息化平台农技咨询数据库的构建. 中国科技成果, 2013(2): 38-39.



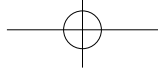
- [8] 温孚江, 宋长青. 农业大数据应用、研究与展望. 农业网络信息, 2017(5): 31-36.
- [9] 海云瑞, 张学俭. 宁夏农业物联网大数据共享及综合应用平台研究与设计. 宁夏农林科技, 2017, 58(9): 51-53.
- [10] 康春鹏, 董春岩, 王文月, 等. 我国农业农村大数据发展应用研究, 2018, 30(6): 100-104.
- [11] 郭雷风. 面向农业领域的大数据关键技术研究. 北京: 中国农业科学院, 2018.
- [12] Perakis K., Lampathaki F., Nikas K., et al. CYBELE—Fostering precision agriculture & livestock farming through secure access to large-scale HPC enabled virtual industrial experimentation environments fostering scalable big data analytics. *Computer Networks*, 2020(168): 107035.
- [13] Shrivastava S., Marshall-Colon A. Big data in agriculture and their analyses. *Encyclopedia of Food Security and Sustainability*, 2019(1): 233-237.
- [14] Sabarina K., Priya N. Lowering data dimensionality in big data for the benefit of precision agriculture. *Procedia Computer Science*, 2015(48): 548-554.
- [15] Huang Yanbo, Chen Zhongxin, Yu Tao, et al. Agricultural remote sensing big data: Management and applications. *Journal of Integrative Agriculture*, 2018, 17(9): 1915-1931.
- [16] 李艳, 刘成龙. 基于 Hadoop 的农业大数据挖掘系统构建. 农业图书情报学刊, 2019(2): 70-71.
- [17] 钱晔, 孙吉红, 张悦. 智能算法在农业大数据中的探究. 北方园艺, 2019(20): 156-161.

Research and development of agricultural big data sharing service platform based on Cloud Computing

Yu Wanmin¹, Fan Beilei^{2*}, Qian Jianping³

(1. Guizhou Province Agriculture Information Center, Guiyang 550001, China; 2. Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China; 3. Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: [Purpose] In order to integrate multi-source, heterogeneous and massive agricultural data, an open sharing platform for agricultural big data is designed and implemented based on cloud computing technology, which is conducive to the full utilization of agricultural data resources. [Method] The platform is designed according to scientific and reasonable standards and specifications, taking agricultural big data resource catalog metadata, big data integration and database building, agricultural big data sharing service as the main line, using Hadoop technology, and based on J2EE platform and component-based development ideas. [Result] The platform includes agricultural big data integration management, agricultural big data sharing service management and metadata management. Agricultural big data integration management subsystem realizes the integration, data quality control and update management of agricultural data obtained from various channels. Agricultural big data sharing service subsystem realizes the directory navigation and query of decentralized, multi-source and heterogeneous agricultural non spatial data query retrieval, and agricultural big data sharing service subsystem is the unified entry for users to query, access and use agricultural big data. The metadata management subsystem realizes the unified description, release management and authority



2020 年 2 月

control of multi-source big data. [**Conclusion**] The platform has been developed and applied in Guizhou Province, which realizes the unified integration, management, sharing and service of agricultural data resources in Guizhou Province, and provides the data foundation for agricultural cloud service and its various applications in Guizhou Province.

Key words: agricultural big data; resource directory; cloud computing platform; data sharing

欢迎订阅《中国农业信息》

《中国农业信息》(双月刊)由农业农村部主管,中国农学会农业信息分会、中国农业科学院农业资源与农业区划研究所共同主办,是我国目前全方位传播和刊载国内外农业遥感/农业信息科学领域的信息获取、处理、分析和应用服务的理论、技术、系统集成、标准规范等方面最新进展和成果,促进学术交流以及农业信息学科关键技术与产品的创新研发、集成推广和应用示范的综合性科学技术期刊。

主要刊登农业遥感、农业传感器、农业信息智能处理、精准农业/智慧农业、农业监测预警与信息服务系统、农业物联网、智能装备与控制、虚拟农业、人工智能、信息技术标准等方向学科热点领域的最新、最重要的理论研究和应用成果。主要栏目有:农业遥感、智慧农业、综合研究、农业信息技术、农业物联网、专题报道等。目前被中国知网(CNKI)、万方数据、中文科技期刊数据库、中国核心期刊(遴选)数据库等多家数据库收录。

《中国农业信息》为国内外公开发行的刊物,开本为16开,彩色四封,读者范围广,影响面大,全国各地邮局均有订阅。每双月25号出版,定价为25.00元/册,150元/年。

邮局汇款

收 款 人:《中国农业信息》编辑部

地 址:北京市海淀区中关村南大街12号中国农科院资源所区划楼319

邮 编:100081

银行汇款

开 户 行:农行北京北下关支行

行 号:103100005063

账 号:11050601040011896

单位名称:中国农业科学院农业资源与农业区划研究所

电 话:(010)82109632

传 真:(010)82109632

E m a i l : nyxbjb@caas.cn

邮发代号:2-733

投稿网址:www.cjarrp.com