

基于云端的农业病虫草害大数据图文数据库的设计与实现

朱静波, 张立平, 董伟*, 李闰枚, 方钰, 孔娟娟 (安徽省农业科学院农业经济与信息研究所, 安徽合肥 230031)

摘要 针对病虫草害大数据的存储分散和展示无系统化的问题, 提出采用 Spark 核心技术搭建大规模集群, 将 HDFS(Hadoop distributed file system) 分布式文件存储系统、MongoDB 数据库和 MySQL 数据库相结合, 集病虫草害信息管理、信息查询、用户管理和数据库维护等功能于一体, 实时更新、展示、存储和管理海量多源异构病虫草害数据, 构建基于云端的农业病虫草害大数据图文数据库信息服务平台, 在提高农业病虫草害的数字化管理、信息共享等方面具有极其重要的意义。

关键词 病虫草害; Spark 框架; 大数据; 图文数据库

中图分类号 S126 文献标识码 A 文章编号 0517-6611(2017)16-0206-04

Design and Implementation of Big Data Graphic Database for Agricultural Diseases, Pests and Weeds Based on Cloud Calculation
ZHU Jing-bo, ZHANG Li-ping, DONG Wei* et al (Institute of Agricultural Economy and Information, Anhui Academy of Agricultural Sciences, Hefei, Anhui 230031)

Abstract Aiming at the problem of big data storage dispersion and display without systematization of diseases, pests and weeds, Spark core technology was used to build large-scale clusters, HDFS distributed file system, MongoDB database were combined with MySQL database to construct big data graphic database system of agricultural diseases, pests and weeds based on cloud technology. The system implemented the functions of information management, information query, user management and database maintenance. The database system had strong compatibility, clear classifications, friendly interface and provide real-time updates, massive data storage and management for heterogeneous multi-source data display and pest and. Also, integrated information service platform was established for the diseases, pests and weeds information, and it was of very important significance for strengthening the digital management, information sharing etc..

Key words Diseases, pests and weeds; Spark framework; Big data; Graphic database

农业病虫草害是我国主要农业灾害之一, 具有范围广、种类多、发生情况复杂等特点, 每年都会产生大量的病虫草害数据信息^[1-2]。随着我国农业信息化、产业化发展进程的推进, 对这些数据信息进行实时采集、传输、管理和分析挖掘处理已经成为促进农业病虫草害防治的一个重要抓手。虽然病虫草害大数据研究的重要性日益凸显, 但目前大部分研究都针对专业的数据挖掘和分析, 而基层农技员、种植大户、农民等主要使用者反而较难接受其数据展示含义^[3]。

目前, 我国关于农业病虫草害大数据展示的研究较少, 主要集中在对农业大数据的采集、分析与处理^[4]。陶佰睿等^[5]于 2016 年提出农业大数据云端系统, 该系统以 Hadoop 集群分布式存储为基础, 融合 MySQL 数据库, 旨在能够实时监测农作物的生长过程, 从而提高生产率; 柴进^[6]选取中国农业科学院提供的农业数据为研究对象, 设计出基于 Hadoop 平台的农业数据挖掘系统。此外, 针对病虫草害数据量少的分析展示系统也无法满足现有信息需求, 多数为部分单一品种数据库的构建, 无法满足用户对病虫草害的认知需求。王长委等^[7]针对 GIS(Geographic information system) 在病虫害数据管理方面的优缺点, 通过 Geodatabase 构建了广东省农作物病虫害空间数据库; 余一鸣^[8]通过多角度整理和分析上海地区的蔬菜病虫害基础图文资料, 搭建了上海蔬菜病虫害知识普及和防治技术咨询的服务平台。针对现有病虫草害数据的大规模、繁结构、多形式和防治技术繁多等突出的大数据特点, 笔者在深入分析作物种类的基础上, 提出采用

Spark 核心技术搭建大规模集群, 使用 HDFS 分布式文件系统、MongoDB 数据库和 MySQL 数据库相结合, 搭建基于云端的农业病虫草害大数据图文数据库系统, 旨在实时更新、展示多源异构病虫草害数据和对相关的数据进行储存和处理。

1 相关概念及关键技术

1.1 农业病虫草害大数据的特点 随着互联网及物联网的发展, 我国病虫草害各类数据资源越来越多, 逐步呈现出数量巨大、结构复杂、形式多样和防治技术繁多等明显的大数据特征, 使得人们无法快速获取、查阅和储存相关病虫草害信息。①病虫草害发生期间产生庞大数据量。现代化农业对病虫草害的监测现已趋于成熟, 各类传感器分布在农业种植区域实时监测农业病虫草害发生前中后期所体现的病征, 以至产生的数据量达到 PB(petabyte) 单位级^[9]。②病虫草害数据种类繁多。我国种植作物的种类已超过千种, 不同作物不同生长时期发生的病虫草害都有可能存在不同, 这使得病虫草害的种类繁多。③病虫草害数据类型多样。数据来源于遥感传感, 视频、图片多媒体, 表格、文本, 特征表现为多源异构。④病虫草害数据存储。需要结合分布式存储、云计算, 实现低成本、低功耗、高可靠的目标。

1.2 基于云端的农业病虫草害大数据关键技术 病虫草害大数据的发展对传统的数据存储、分类处理、分析展示技术体系提出巨大挑战, 必须要在采集数据、标准设置、分析处理、系统展现等方面进行全新的技术升级。针对数据存储, 在多个商用机器上运用 Spark 搭建大规模集群, 并使用 HDFS 分布式文件系统、MongoDB 数据库; 针对数据展示, 依据作物基本情况将病虫草害进行分类, 分类别展示作物病虫草害数据^[10]。

目前, 传统关系型数据库应用较为成熟的有 Oracle、

基金项目 安徽省农业科学院院长青年创新基金项目(17B1422)。
作者简介 朱静波(1980—), 男, 安徽合肥人, 研究实习员, 硕士, 从事农业信息方面的研究。* 通讯作者, 助理研究员, 硕士, 从事植保信息化方面的研究。

收稿日期 2017-04-14

MySQL 数据库,然而已有的病虫害数据库对分布式处理不能提供较为良好的支持,对多用户的高并发访问和计算有所欠缺,共享性和扩展性不强。因此,将多源异构数据存储在高性能数据库 MongoDB 上,放置于 Spark 云端,发挥其扩展性快速、灵活处理相关数据,从而构建基于云端的农业病虫害大数据图文数据库系统以供用户使用。系统结构如图 1 所示。

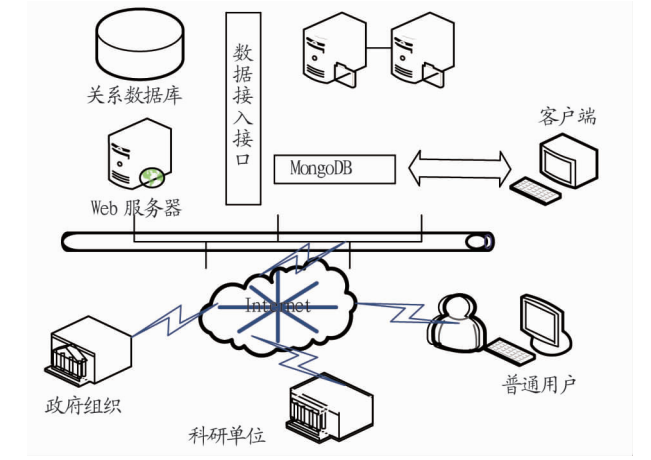


图 1 系统结构
Fig.1 System structure

2 病虫害大数据获取与分类

2.1 病虫害大数据的获取 多源性、异构性是病虫害大数据的主要特征,是数据获取的难题之一。针对该特性,选取病虫害大数据的结构化数据和部分非结构化数据进行存储和展示。结构化数据是经过专业化、系统化处理后的数据,可通过建立索引存储在数据库中,例如现有专业病虫害数据库、农业信息网站等。

病虫害大数据的收集主要通过以下 2 个方式:①高校和相关农业科研单位。安徽省农业科学院农业经济与信息研究所已在病虫害信息资源采集与整理领域开展了近 20 年的工作,采集各类具有自主知识产权的病虫害害图像 50 余万幅,整理文字逾 1 000 万字,出版病虫害科普读物近百册,且与多家科研院所保持长期合作关系,联合开展相关的数据收集与整理。数据获取方式主要包括科研调查、实地采集、试验监测、专家咨询等。这些数据是该数据库的主要数据源,通过对获取的数据进行结构化整理、分析、审核,最后按照分类模式存储数据,从而实现了数据的共享、可视化和应用分析。②农业信息网站和相关数据库。在信息化快速发展的社会,农业信息网站和相关数据库为数据获取奠定了基础。例如,中国农业信息网以及各地的农业信息网,提供丰富而又全面的农业信息资讯和具有当地特色的病虫害害图文数据;其他涉农数据库,如农业科技信息数据库、农业科技期刊全文数据等,也均可提供丰富的数据。此外,一些综合性数据库也提供了大量的病虫害害数据,如中国知网、万方数据、维普等,只要键入病虫害害关键词,就可以获取相关论文、成果、专利等数据信息。

2.2 病虫害大数据的分类 我国是农业大国,农作物种

类繁多,根据用户使用习惯,病虫害数据的分类沿袭了农作物的分类,分为农作物病虫害、蔬菜病虫害和果树病虫害。其中,农作物主要包括禾谷类作物、薯类作物、绿肥作物、豆类作物、经济作物、热带亚热带作物等;蔬菜包括白菜类、甘蓝类、绿叶菜类、茄果类、豆类、瓜类和多年生菜类等;果树则主要包括仁果类、核果类、浆果类、坚果类、杂果类、柑桔类和热带及亚热带果类。每个二级分类下又对应着不同的产品,如禾谷类主要包括水稻、小麦、大麦、玉米、高粱、粟等。鉴于杂草的防治方法与其生物学分类地位相关性较高,笔者设计的图文数据库对杂草的分类也按生物学分类进行划分,主要包括孢子植物、单子叶植物和双子叶植物。

3 数据存储层及数据库设计

3.1 数据存储层架构和部署 主要使用 MySQL 数据库存储无需改变的数据,采用 Spark 框架构建分布式存储管理平台,对实时传感数据进行存储管理,实现数据存储的高可靠性、高扩展性。随着农业作物种类的增长,农业病虫害害数据量呈爆炸式上升,现有的数据库系统已无法满足海量数据的存储与计算要求,因此笔者设计了实现基于 Spark 云计算平台的病虫害害图文数据库,实现海量数据的高效安全的存储和精确的分类信息展示。

Apache Spark 是近年来发展较快的分布式并行数据处理框架,是专为大规模数据处理而设计的快速通用的计算引擎。Spark 是一种与 Hadoop 相似的开源集群计算环境,但 Spark 使用内存缓存机制,它不仅能够提供交互式查询外,而且优化了迭代算法的性能,提高计算效率。此外,Spark 还增加了流数据处理、图数据处理等更为高级的数据处理能力,可用于构建大型、低延迟的数据分析应用程序,其生态系统如图 2 所示。

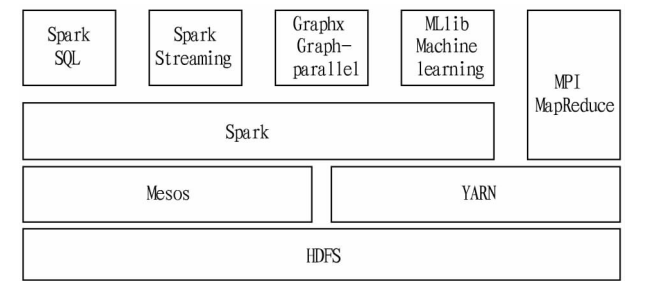


图 2 Spark 的生态系统
Fig.2 Ecological system of Spark

在搭建服务器的集群环境时,设置 2 台服务器分别作为主节点和备用主节点,每台服务器上安装 Spark;另设 3 台服务器作为子节点,配置为 4 GB 内存、1 T 硬盘。系统部署模型如图 3 所示。

3.2 数据库内容设计 针对现有农业病虫害害特性与防治的研究,农业病虫害害图文数据库对其进一步分析处理,主要包含基本描述数据、为害症状数据、发生因素数据、病原物数据、侵染循环数据、防治方法数据和图册数据等信息。其中,基本描述数据是指描述不同作物病虫害害的基本属性信息,包括病虫害害的中文名称、学名、异名和简要介绍等;为害症状数据是描述病虫害害的种类和其对应症状信息,包括

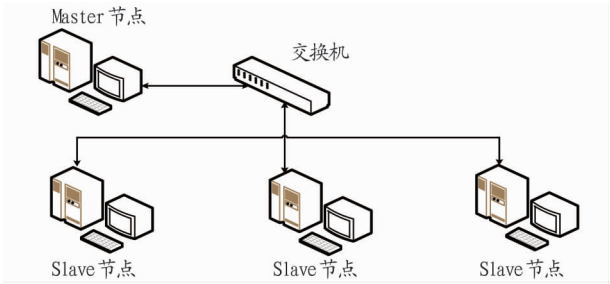


图3 系统部署模型

Fig.3 The deployment model of the system

病虫害害的种类、叶片病斑描述、植物损害等;发生因素数据是指描述病虫害害发生密切相关的生物学特征,包括发生期、温度、湿度以及其他相关天气因素;病原物数据是指致使发病的病原物的基本生理信息,包括病原物名称、属类、性状、生长环境和周期等;侵染循环数据则是描述病虫害害侵入植物的过程,主要包括初侵染、侵入和传播;防治方法数据主要描述针对病征所采取的防治方法,包括基本防治、种子预处理、药剂治理和生物治理;图册数据则是展示病害、害虫和杂草的外观形态等。

3.3 数据库结构设计 为了持久化存储病虫害害数据,选用基于分布式文件存储的、可为 WEB 应用提供可扩展的高性能数据存储,可存储复杂数据类型的 MongoDB 数据库,并将其与 MySQL 数据库相结合,为数据存储提供技术支撑。

数据库包含类别信息、类型信息、标签信息、病虫害害信息和图片信息 5 张数据表以及 1 个病虫害害图库。除病害外,同一种作物上可以发生多种害虫和杂草,同一种害虫和杂草也可以发生在多种作物上,所以害虫和杂草与作物的联系皆是一对多的关系。通过以上对病虫害害相关内容的深入研究与分析,在 MongoDB 中存储时,病虫害害信息以描述其基本信息,其存储逻辑如表 1 所示。图片信息表则存储病虫害害病征图片的名称,通过与图库中的图片构建索引进行查找,其存储逻辑如表 2 所示。病虫害害图库中对应图片信息表存储各个病虫害害的病征图片,使用时将图片转换成图片流从本地上传至 HDFS,从而被 Web 系统所利用。

表 1 病虫害害信息的存储逻辑

Table 1 Storage logic of disease, pest and weed information

字段名称 Field name	数据类型 Data type	备注 Description
ID	int	病虫害害 ID,主键约束
Type_ID	int	类型 ID,与类型信息表的 ID 字段进行外键约束
Name	varchar(50)	名称
English_name	varchar(50)	英文名称
Alias	varchar(50)	异名
Description	varchar(500)	基本描述
Harm_symptom	varchar(500)	为害症状
Pathogen	varchar(500)	病原物
Infection_cycle	varchar(500)	侵染循环
Occur_factor	varchar(500)	发生因素
Control_method	varchar(500)	防治方法
Update_time	datetime	添加时间

表 2 图片信息的存储逻辑

Table 2 Storage logic of picture information

字段名称 Field name	数据类型 Data type	备注 Description
ID	int	ID 号,主键约束
Photo_name	varchar(50)	图片名称及路径,对应图库中的图片
Disease_ID	int	病虫害害 ID,与病虫害害信息表的 ID 字段进行外键约束

4 系统设计与实现

4.1 系统设计 基于云端的农业病虫害害大数据图文数据库系统分为前台系统和后台系统:前台系统主要提供基本描述、为害症状、发生因素、病原物、侵染循环、防治方法、图册等七大模块,充分展示服务器端存储的病虫害害数据库;后台系统则主要提供系统管理、病虫害害管理和类别管理 3 个模块,从而实现客户端对服务器端数据的添加与更新。系统总体设计如图 4 所示。

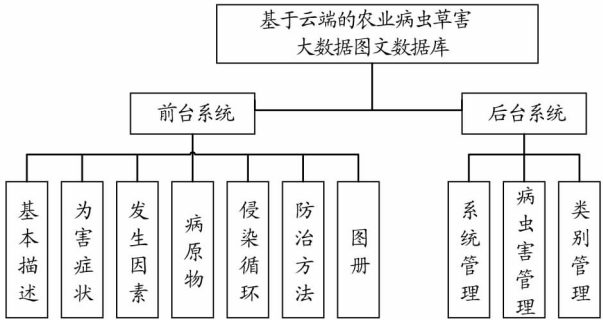


图 4 系统总体设计

Fig.4 Overall design of the system

4.2 系统应用 农业病虫害害图文数据库系统目前已经正式投入使用。图 5 为数据库中水稻病害稻瘟病的图文界面,用户可查看稻瘟病英文名、异名、简介、为害症状、发生因素、病原物、侵染循环、防治方法和相关病征图片等。



图 5 稻瘟病信息界面

Fig.5 The interface of rice blast information

农业病虫害害图文数据库后台系统则分为系统管理和业务操作,其中系统管理主要提供管理员信息的设置;业务操作则主要包括病虫害害管理和类别管理,病虫害害管理页面可实现病虫害害信息的添加和已有信息的更改,类别管理页面

可实现农业种类的添加与编辑。后台管理系统提供了丰富的数据管理功能,能够实现对病虫害大数据的快速更新与展示。后台管理界面如图 6 所示。

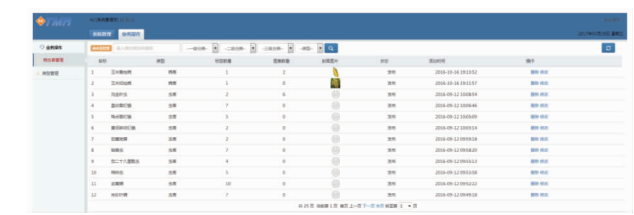


图 6 后台管理界面

Fig.6 The interface of backstage management

5 结语

笔者在深入研究农业病虫害大数据的基础上,对各类病虫害相关信息进行了总结,结合数据库系统功能与结构的研究,采用 Spark 核心技术搭建分布式集群,利用 HDFS 分布式文件系统、MongoDB 和 MySQL,设计实现基于云端的农业病虫害大数据图文数据库系统,实现海量病虫害数据的实时获取、更新以及高效的存储管理。笔者采用多级分类的形式,解决单一品种数据库无法满足现有对病虫害认知需求的问题。该系统在应用与推广后,验证了其可行性,并具有以下优点:界面清晰、操作便捷;采用HTML5自适应技

(上接第 197 页)
动与之密切相关。

(3)微观上,即从指标层因素的障碍度出发,在即将到来的年份里影响岳阳市土地利用效益的障碍因子主要集中在造林面积、人均城乡居民储蓄余额、每万人拥有在校大学生人数、人均社会消费品零售总额和每万人拥有公交车辆数。

4.2 讨论 该研究着眼于系统,由此出发构建岳阳市土地利用效益评价体系 and 障碍度模型,探索了岳阳市土地利用效益和影响其的障碍因子,可以为今后岳阳市土地利用效益的提高提供决策参考依据。当然,该研究亦有诸多不足,在数据上,未能获取最新数据,对城市土地利用体系认知不充分,在评价土地利用强度上缺乏指标,在土地管理上也并未设置评价指标,这些皆对研究土地利用效益和障碍分析构成局限,并且也未对各个评价指标做关联度分析,各指标之间的相互关系无法精确辨识,从而导致系统分析的不完整性,有待今后更进一步研究。

参考文献

[1] 彭建,蒋依依,李正国,等.快速城市化地区土地利用效益评价:以南京市江宁区为例[J].长江流域资源与环境,2005,14(3):304-309.
[2] 汪洋,代立,周颖.基于可持续性视角的城市土地利用模糊逻辑评价与诊断:以武汉市为例[J].中国土地科学,2016,30(4):61-69.

术,兼容互联网各类平台,扩展性好;农业种类分类详细,即使是普通农户依然可以快捷获取病虫害草害相关信息。此外,该数据库系统的设计与实现,为病虫害草害信息建立了集成的信息服务平台,在规范农业病虫害草害的数字化管理、促进信息共享、为农业病虫害草害研究和防治提供理论支撑等方面发挥了积极的作用。

参考文献

[1] 籍延宝.农业主要病虫害监测预警系统通用平台的开发及初步应用[D].北京:中国农业大学,2014.
[2] 杨怀文.我国农业病虫害生物防治应用研究进展[J].科技导报,2007(7):56-60.
[3] 谢润梅.农业大数据的获取与利用[J].安徽农业科学,2015,43(30):383-385.
[4] 华雪琦,孙明喆,赵慧彤,等.农业物联网数据共享系统设计与研发[J].农业技术与装备,2016(1):41-43,46.
[5] 陶佰睿,李春辉,苗凤娟,等.基于云端的农业大数据采集与管理系统浅析[J].工业和信息化教育,2016(9):81-89.
[6] 柴进.基于Hadoop的农业数据挖掘系统的研究与实现[D].北京:北京工业大学,2015.
[7] 王长委,胡月明,张俊平.面向对象的病虫害数据库研究与建立[C]//中国地理信息系统协会第四次会员代表大会暨第十一届年会论文集.北京:中国地理信息系统协会,2007:703-710.
[8] 余一鸣.上海蔬菜病虫害及其防治技术咨询服务平台的构建与应用[D].上海:上海师范大学,2015.
[9] 倪冬平.面向农作物病虫害草害数据的XML转换和检索方法研究与应用[D].合肥:安徽农业大学,2012.
[10] 郝进义.数据库设计规范及设计技巧研究[J].计算机光盘软件与应用,2012(12):176-177.

[3] 范辉,刘卫东,周颖.基于结构-功能关系的城市土地集约利用评价:以武汉市中心城区为例[J].经济地理,2013,33(10):145-150,187.
[4] 朱一中,曹裕.基于PSR模型的广东省城市土地集约利用空间差异分析[J].经济地理,2011,31(8):1375-1380.
[5] 宫继萍,石培基,潘竟虎.基于DEA方法的兰州市城市土地利用经济效益分析[J].水土保持通报,2011,31(5):163-166.
[6] 鲁春阳,文枫,杨庆媛,等.基于改进TOPSIS法的城市土地利用绩效评价及障碍因子诊断:以重庆市为例[J].资源科学,2011,33(3):535-541.
[7] 刘爱,宋成舜,翟文侠,等.基于微观尺度的城市建设用地集约利用评价:以西宁市为例[J].资源开发与市场,2011,27(9):812-815.
[8] 刘畅,师学义,梁旭琴,等.基于物元模型的资源型城市土地利用效益动态评价研究[J].水土保持研究,2015,22(4):122-126,131.
[9] 张荣天,焦华富.长江经济带城市土地利用效率格局演变及驱动机制研究[J].长江流域资源与环境,2015,24(3):387-394.
[10] 谭荣,曲福田.从土地利用效益看农业布局 and 结构调整:以江省为例[J].经济地理,2006,26(6):1033-1036.
[11] 张春慧,郑荣宝,陈美招,等.主体功能区划基础上的珠江三角洲土地利用效益评价研究[J].资源开发与市场,2016,32(1):46-50.
[12] 胡永宏.对TOPSIS法用于综合评价的改进[J].数学的实践与认识,2002,32(4):572-575.
[13] 张俊凤,徐梦洁.城市扩张用地效益评价与耦合关系研究:以南京市为例[J].南京农业大学学报(社会科学版),2010,10(3):63-69.
[14] 陶江,吴世新,董雯.天山北坡经济带土地利用效益评价[J].干旱区地理,2009,32(6):985-990.
[15] 李娅婷,张妍.北京农业循环经济发展评价研究[J].环境科学与管理,2009,34(1):109-112.