基于 Hadoop 农业大数据管理平台的设计①

文 燕

(成都农业科技职业学院, 成都 611130)

摘 要: 信息技术的高速发展使得每天的数据量以 TB 级速度暴增, 如何有效利用和管理这些爆炸式增长的大数据呢？是当前亟待处理的问题. 大数据已经渗透到包括农业领域在内的各个领域, 随着农业信息化建设以及物联网技术在农业生产中的应用, 产生了海量的农业大数据待存储、管理和处理. 本文以成都农业科技职业学院彭州葛仙山农业示范基地的农业信息化建设为背景, 根据农业物联网和信息化建设要求, 构建高性能基于 Hadoop 农业大数据管理的平台, 实现农业大数据的安全可靠存储、智能管理与应用, 最终达到对农业生产的智能预警、智能决策和智能分析的目的, 并为农户提供专业的指导. 为我国进入精细化种植、精准化控制、可视化管理、智能化决策的智慧农业时代奠定基础.

关键词: 农业大数据; Hadoop; Map/Reduce; HDFS; 智慧农业

Design of the Management Platform for Agriculture Big Data Based on Hadoop

WEN Yan

(Chengdu Agricultural Science and Technology Vocational College, Chengdu 611130, China)

Abstract: The rapid development of modern information technology makes the amount of every day data increase at the speed of TB, how to effectively use and manage the big data with explosive growth? It is a problem need to be solved urgently. Big data has penetrated into various fields including agriculture, with the agricultural informatization construction and the application of Internet technology in agricultural production, resulting in a large amount of agricultural data to be stored, managed and processed. Based on the background of the construction of agricultural informatization in Pengzhou Gexian mountain agricultural demonstration base of Chengdu agricultural science and technology vocational college, according to the demand of informatization construction of agriculture, we build high performance agricultural big data management platform based on Hadoop, realizing the agricultural big data safe and reliable storage, intelligent management and application. And ultimately we achieve the purpose of early intelligent warning of agricultural production, intelligent decision-making and intelligent analysis, providing professional guidance to farmers. This lays the foundation for China to enter the intelligent agricultural era based on fine planting, precise control, visual management and intelligent decision-making.

Key words: agriculture big data; Hadoop; Map/Reduce; HDFS; wisdom agriculture

1 引言

“大数据”的相关概念, 早在 1980 年由阿尔文.托夫勒出版的《第三次浪潮》[1]中已经提出. 随着物联网、云计算等技术的迅猛发展, 大数据再次吸引了人类的眼球. 2015 年中央一号文件再次聚焦农业, 主题为进一步深化农村改革加快推进农业现代化、信息化, 这也是中央连续 12 年一号文件关注农业. 农业作为一个国家的基础产业, 也紧随着时代的步伐, 加强现代化大农业发展, 加快科技创新, 实施重大农业科技创新, 积极开展应用基础和前沿高技术领域自主创新[2], 传统的农业生产方式应向数据驱动的智慧化生产方式转变, 标志着进入农业大数据时代.

① 基金项目:四川省教育厅 2016 年四川省高校人文社科学重点研究基地科研项目(TCCSJY-2016-C16);成都农业科技职业学院科研项目(成农院[2016]1-24)

收稿时间:2016-08-10;收到修改稿时间:2016-09-23 [doi:10.15888/j.cnki.csa.005737]

随着农业信息化的不断推进, 在长期的研究和实践过程中, 通过观察、测量、实验等方式积累了大量的对农业生产经营过程具有实际指导意义的农业数据, 而且这些数据还在呈几何级飞速增长. 这些飞速增长的数据形成了农业大数据, 它是由结构化和非结构化的数据组成, 它涉及到农业生产经营过程中的方方面面, 比如育种、耕种、收割等[3]. 农业的快速发展和农业物联网的应用, 非结构化数据在农业数据的比重的逐渐上升, 将很快会远远超过农业数据中结构化数据.

如何管理和利用蕴含大量的价值的数据, 是人类亟待解决的问题. 目前, 在管理和处理农业大数据方面存在一些凸显的问题:

(1) 各级农业部门信息孤立, 各自为阵, 主要以结构化的关系型数据存储方式来存储.

(2) 农业大数据具有自身的特点, 如: 土壤类型众多, 作物品种复杂, 病虫害发生频繁且症状不断变化, 肥水、气候相互之间的关系和影响, 就使得关于它们的数据库与知识库具有大型、多维、动态、不完全、不确定等特征[3].

(3) 各级部门对农业生产和经营过程中采集的数据的重视程度不够. 对于花费了大量的人力、物力和财力建立起来的农业大棚, 采集来的数据不够全面, 数据类型过少, 并且对于采集来的数据也没有得到及时的处理与存储和有效的管理, 就更别说有效的利用, 因此农业的智能化程度不高[4].

(4) 对于农业大数据集中管理和利用效率不高.

(5) 在农业大数据的存储和价值挖掘方面, 对传统存储方式和计算平台已远远不能够满足农业大数据的处理需要.

随着农业信息技术的普及, 成都农业科技职业学院依托智能化农业大大背景, 在彭州葛仙山农业示范试验基地进行了一系列信息化和智能化建设, 为了方便对彭州葛仙山农业产业示范园所采集的农业大数据更好的管理和更高效的利用, 真正发挥示范基地的作用, 因此, 迫切的要构建针对彭州葛仙山农业产业示范基地的海量数据的存储和管理的大数据处理平台. 针对农业大数据自身的特征, 搭建基于Hadoop的农业大数据管理应用平台, 对采集的结构化和非结构化的农业大数据进行并行处理, 挖掘出有价值的数据为农业生产和科研服务, 充分发挥大数据在智能化、现代化农业产业中的作用.

2 Hadoop平台简介

Hadoop 平台是由 Apache 开发的一个运行在廉价机器上的开放式、可扩展的分布式计算框架, 是一种底层细节透明的分布式集群系统架构, 即用户在不了解底层是实现的基础上, 可以根据自身需求, 通过函数编程和操作接口进行应用开发的分布式系统[5].

2.1 Hadoop 分布式结构模型

Hadoop 分布式数据处理框架包括 HDFS(分布式文件系统)和 Map/Reduce(分布式处理模型)两个核心的引擎, 还包括了 HBase(非关系型数据库)、Hive 等大量的组件, 其结构模型如图 1 所示.

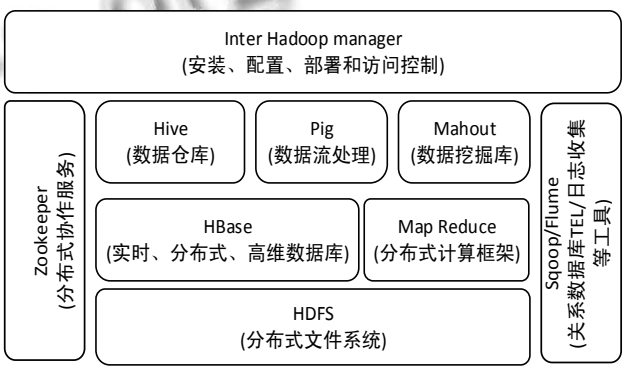

图 1 Hadoop 分布式结构模型

2.2 HDFS(分布式文件系统)

HDFS(Hadoop Distributed File System)是 Hadoop 分布式结构中非常的核心部件, 主要是对数据进行存储和管理, 优点是容错性高、高吞吐量、有一定的硬件故障检测能力. 因此, HDFS 即使部署在廉价的硬件平台上, 都能够通过流式数据访问的方式提供高吞吐量的数据访问能力, 从而提高整个应用系统的性能, 对于海量的农业大数据的应用系统非常适合.

HDFS 采用主从架构(Master/Salve)的结构模式, 包括一个控制节点负责管理和存储 Hadoop 系统数据

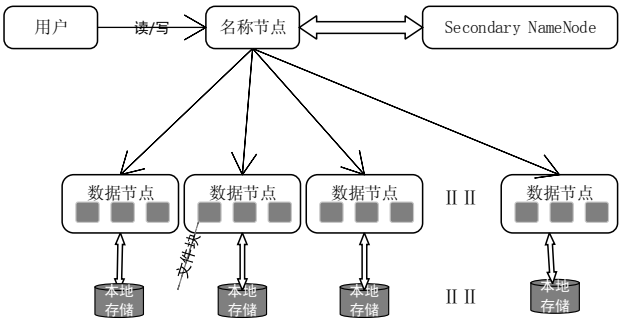

图 2 HDFS 架构模型

信息的位置和名称空间, 处理客户端的请求, 一般定义为 NameNode(名称节点); 一定数量的数据节点, 主要负责存储实际数据, 告诉主节点存储信息, 一般定义为 DataNode(数据节点), 需要存储的切分文件为: Client. HDFS 架构如图 2 所示.

2.3 Map/Reduce 并行计算框架

Map/Reduce 是一种新的分布式环境下的并行计算模型, 由谷歌实验室于 2004 年发表的论文中提出的, 主要适用于大于 1TB 的大规模数据集计算分析. Map/Reduce 是 Map 函数和 Reduce 函数两个核心操作组成, 其中 Map 函数对 Client 传来的热切文件按照一定的规则映射成一组相关的中间文件; Reduce 函数则是对 Map 函数传来的中间文件按照规约进行合并或缩减, 得到最终的结果. Map/Reduce 架构模型如图 3 所示. 该模型由 Job Tracker 总调度, 把每一个任务分配给 Task Tracker 执行, 运行在 HDFS 上的各数据节点的 Task Tracker.

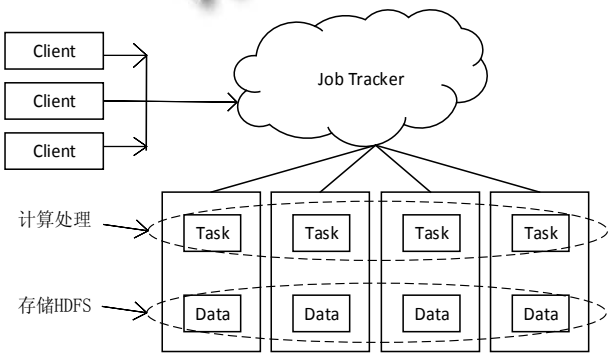

图 3　Map/Reduce 架构模型

2.4 HBase 分布式数据库

HBase 是 Hadoop 分布式结构中的一个重要组件, 其理论基础来源于名为《Bigtable: 一个结构化的数据库分布式系统》Google 的论文, 是当下最为流行的分布式数据库系统. 与传统的关系型数据库相比, HBase 最大的不同之处存储格式采用的是列式存储格式, 对于非结构化的数据库存储特别适合列式存储格. HBase 底层为 HDFS 分布式文件系统, 使用 ZooKeeper 实现分布式协同机制, 利用 Hadoop 的 Map/Reduce 组建来处理海量数据存储在数据库.

Hadoop 分布式架构平台对于解决异构农业大数据的存储管理, 挖掘有效资源提供非常重要的开源架构平台, 主要优点: 可靠性高、可扩展性好、高效性、容错性高、低成本.

3　基于Hadoop农业大数据平台的设计

针对现有农业数据处理平台的数据处理和存储存在不足. 一是现有农业数据处理平台采用的集中式数据库架构, 随着数据的不断增多, 数据库的性能会受到严峻的影响, 这也是集中式数据库成为整个平台架构的瓶颈, 而对于海量的非结构化农业大数据更是无法解决. 二是针对海量非结构化的农业大数据, 现有解决平台的顺序计算耗时长, 不能够满足农业大数据时间要求. 因此, 要解决农业大数据的计算处理、存储和挖掘问题, 需要在现有农业大数据管理平台的基础上构建以 Hadoop 为计算处理中心, HDFS 和 HBase 为数据存储中心的农业大数据管理平台. 主要借助于 Hadoop 分布式并行计算的数据处理能力以及 HDFS 和 HBase 分布式大数据存储能力, 为农业大数据的处理和存储提供了保障.

3.1 Hadoop 农业大数据管理平台架构

通过对成都农业科技职业学院彭州葛仙山农业信息化示范园的实际情况分析, 结合 Hadoop 的体系结构的研究, 提出如图 4 所示基于 Hadoop 农业大数据管理平台架构. 该农业大数据平台从底层往上依次农业大数据采集层、数据存储中心、计算处理中心、交互层与智能农业终端应用系统. 各部分之间通过网络通信和数据传输保证整个系统正常运行.

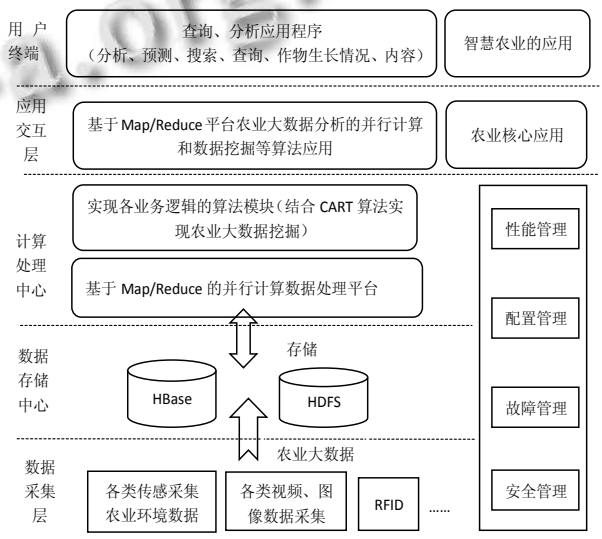

图 4　Hadoop 农业大数据管理平台架构

下面对平台各部分的功能进行介绍:

① 数据采集层

已经建成的农业大数据采集网络, 利用各类智能终端采集设备传感器、RFID 和摄像头等, 采集农业生长环境所有的各类环境参数信息、图片和视频信息, 这些信息构成了农业大数据的主要数据来源, 通过网络传输到数据中心.

② 统一的农业大数据存储中心

数据存储中心, 是对农业大数据的管理和存储. 各级农业部门都拥有自己的数据存储中心, 而这些数据存储中心部署在不同的地域, 因此这些数据中心之间不能够统一数据存储格式, 也无法实现信息的共享, 容易形成信息孤岛, 这使得农业数据的计算、处理和价值挖掘不便. 针对于农业大数据具有结构化和非结构化的特点, 该数据存储中心是采用统一农业数据存储中心, 以 Hadoop 中的 HBase 分布式数据库和 HDFS 分布式文件系统为数据管理框架, 不仅可以为上层提供并行的数据访问, 还能够提供高效、安全和易扩展的存储服务. 当系统现有存储能力达到一定极值时, 能够便捷的扩充新的存储节点, 新增存储节点后不会影响原有的数据存储. 与此同时, 为保证数据存储的安全, 该数据管理框架还具有良好的副本机制, 即当存储节点上的数据出现存储异常时, 通过副本机制将数据转移到其他节点.

③ 计算处理中心

计算处理中心是整个系统的核心部分, 为用户提供动态的资源控制、带宽分配、程序开发运行环境, 实现各业务逻辑的功能, 为系统数据处理和数据挖掘提供基础的计算模型, 并为上层提供任务调度模块. 该平台中的计算处理框架以 Hadoop 中的 Map/Reduce 并行计算数据处理平台为基础, 结合 CART 算法实现对农业数据的价值进行挖掘.

④ 农业示范应用

农业示范应用是对该平台的整合和各功能的完备性、正确性的有效验证. 该应用系统平台主要有农业核心应用、农业数据挖掘和智慧农业的应用三类. 农业核心应用主要是基于 Map/Reduce 并行计算框架实现, 包括作物病虫害检查算法、病虫害诊断算法、作物生长情况的分析算法等对原始农业数据进行快速处理的一系列相关算法, 并将处理结果进行存储处理, 以便对事实数据进行查询. 智慧农业的基于农业核心应用中的计算结果, 面向用户需求而设计的, 包括农业数据查询、分析、统计、预测、智能控制、搜索等功能的一系列农业市场的应用.

3.2 农业大数据存储中心解决方案

各级部门数据存储中心通过各种类型传感器、RFID 和视频采集等采集手段获取海量农业数据, 这些数据以不同形式和结构存储在不同地理位置的数据库中. 数据存储中心对分散数据源和异构数据进行有机整合, 并对存储在不同系统的农业原始数据进行高效管理、有效组织和存储, 再通过大数据处理技术解决数据计算的问题.

3.2.1 农业大数据存储中心架构

农业大数据存储中心对于不同地域的各级农业数据中心进行统一组织和管理. 平台通过创建服务实例的方式管理各级数据中心, 每个服务实例对应一个各级的分数据中心, 服务实例记录了原始数据存储中心的地址以及访问的权限等信息, 以及各个数据中心所使用的数据库类型、中心地址、数据库名称、表名称、用户名、登陆密码、访问权限等内容, 从而实现数据中心的资源共享和统一管理. 如用户需要对某个数据中心的数据进行访问时, 只需要通过 Hadoop 平台的中央查询集群中的服务实例就可以查询到对应数据中心的数据. 各数据中心节点和 Hadoop 集群分布式架构如图 5 所示. 分布式集群架构的底层除了部署 Hadoop 集群和 HBase 集群外, 还有 Hadoop 分布式结构模型中的一系列的子项目 Sqoop、Hive 等.

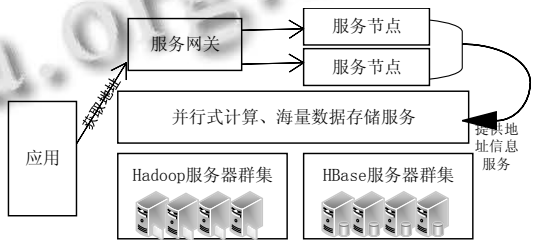

图 5 农业大数据存储中心集群架构

3.2.2 农业大数据并行集群整合服务实例

Hadoop 结构和 HBase 架构都采用的 Master/Slave 结构, 其中 Hadoop 架构是由负责 Map/Reduce 任务调度的 Job Tracker 和负责 HDFS 数据管理节点的 Name Node 构成 Master; HBase 框架中的 Master 由 HMaster 组件构成. 整个集群能否正常运行 Master 起着决定性的作用, 较的好稳定性, 因此对外只提供一个地址服务信息, 即 Master 主机所在地址.

针对农业大数据的并行存储, 各级各地的农业数据存储中心通过创建一个并行集群整合服务实例来实现与 Hadoop 中心访问存储, 实例创建流程如下所示:

1) 用户通过应用命令行终端向服务网关发起创建服务实例的请求.

2) 服务网关接到服务实例创建请求后, 根据平台系统中每个服务节点当前的资源利用情况查找出最优节点, 并通知其创建服务实例.

3) 服务节点在接到服务实例的创建请求后, 记录 Hadoop、HBase 集群地址并记录在对应的服务实例中, 向服务网关返回服务实例创建成功的消息.

4) 服务网关在得到服务实例创建成功的消息后, 在数据库中记录服务实例与服务节点的相关信息, 用于后续与应用的绑定.

客户端应用通过绑定服务实例后, 即可获取分布式集群地址, 与集群进行通信. 通过开放接口输入相关数据, 即可完成 HBase 数据库中表的相关操作以及获取分布式运算和存储环境, 而无需再访问服务数据节点.

3.3 农业大数据计算处理中心的设计

基于 Hadoop 的农业大数据管理平台是一个功能足够强大、便捷、快速的大数据处理平台, 整个平台从数据采集、加工、处理分析、存储、运营和维护提供一条龙服务, 终端用户无须知晓或关注底层如何实现和运维.

3.3.1 基于 Map/Reduce 并行计算框架农业大数据计算处理中心的设计

在 Hadoop 农业大数据平台的数据计算处理中心以 Map/Reduce 并行计算框架作为基础框架, 在基础框架上移植各种算法, 可以实现各种业务逻辑, 以此来满足平台大规模数据集的计算速度和进行数据挖掘. 计算处理中心的结构如图 6 所示.

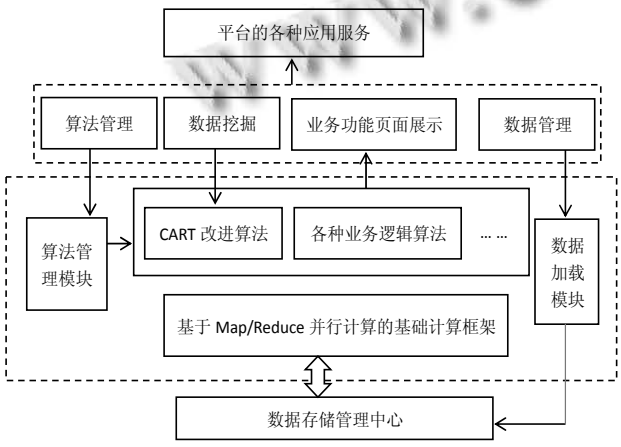

图 6　计算处理中心结构

Map/Reduce 并行计算数据处理是各种数据处理和挖掘算法应用在平台可在不知底层处理细节的情况下, 提供简易交互接口, 实现多种并行计算. 有很好的伸缩性和扩展性, 当系统某一计算节点崩溃时, 该计算框架会自动将崩溃节点的任务分配给其他计算节点; 在平台并行计算处理能力不足的情况下, 可以便捷的增加计算节点, 从而提高整个平台的计算能力.

3.3.2 基于 Map/Reduce 农业数据挖掘解决方案

针对大数据具有海量性, 多样性, 不规则性等特征, 而对于特殊的农业领域的大数据来源于农作物的从生产到餐桌的整个过程, 由于这些数据有类型众多的土壤、品种复杂的农作物、频发的病虫害、不确定的气候等诸多影响因素, 因此, 采集到的相关农业大数据具有不确定、不完全(数据随机噪音)和稀疏性(数据的实用价值不高)等特征. 要让农业大数据在农业生产过程中的起到智能预警、智能决策、智能分析的作用, 需要对农业大数据进行精准、高效的数据挖掘. 对农业领域的数据常见的挖掘主要有: 相关性分析、分类描述、聚类分析、偏差分析等, 而在实际应用中最多的就是数据的分类. 针对农业大数据的特征, 一般选择分类回归树 CART(Classification and Regression Trees)算法.

CART 算法是以统计学为理论基础, 采用的非参数方法, 以典型的二叉树结构为决策树, 即由一个根节点和若干属性节点、叶结点组成, 其分类结果易于理解和掌握. 首先所有的样本集都在根节点内, 然后按照一定的分割方法, 根节点被分割成两个子节点, 样本集也被分割到两个子节点内, 在相同的分割规则下, 递归的对子节点进行分割, 直到不可以再分割为止[12].

基于 Map Reduce 框架农业核心应用以及提取价值数据, 关键在于各种算法的应用, 当然有一些算法如果直接移植到 Map/Reduce 分布式计算框架, 是没有办法完成, 则需要对某些算法做一定的改进. 例如要进行数据的价值挖掘所使用的 CART 算法, 需要改进后才能够将 CART 算法移植到 Map/Reduce 分布式计算框架.

CART 算法本是为解决串行运算问题而设计的, 因具有其特殊性, 在此根据农业大数据的特性可以将 CART 算法并行设计, 基于 Hadoop 平台的农业数据挖掘 CART 算法的并行化设计如下:

1) 计算各个属性 Gini 指数(是判断最佳分裂属性的度量)时的并行, 属性的并行可以通过 Hadoop 中 Map 阶段对定义 Partitioner 来实现, 因为只有相同节点上的相同属性表才会被分发到同一个 Reducer 进行处理.

2) 构建决策树时节点的并行, 从属性的并行设计可知, 同一个节点的所有属性表是一个整体, 一起分割的, 节点分割完成后属性表则会附在新的节点上, 并继续进行分割. 而处在同一层节点之间的产生是不存在相互关联的, 由此在构造决策树时可以对位于树的同一层的所有节点进行并行处理.

3) 排序的并行, 在 Hadoop 平台中, Map/Reduce 在每次分发数据时都会对其进行排序, CART 算法对连续值进行预排序处理, 相邻两个属性值的中间点作为计算 Gini 指数值, 计算时先判断连续性, 再根据属性值的大小进行排序. 对于农业大数据而言, 数据连续值的分布情况以及排序算法的选择对数据挖掘的最终效果会产生很大的影响, 在通过 CART 算法并行设计和改进后, 使其成为并行的算法再结合Hadoop中的Map/Reduce 并行计算框架并行化实现, 使得整个基于Hadoop 的农业大数据平台良好的并行化, 具有较高的数据处理和数据挖掘的能力, 系统的性能也能发挥极致.

4 总结展望

本文对 Hadoop 分布式架构以及其两个核心的引擎 HDFS(分布式文件系统)和 Map/Reduce(分布式处理模型)、HBase 进行详细的分析研究, 提出了 Hadoop 分布式架构大数据平台. 结合成都农业科技职业学院彭州葛仙山示范园实际情况, 对农业大数据的特点进行分析研究, 针对现有农业大数据在存储和处理过程中存在具体问题, 构建出高性能的基于Hadoop农业大数据管理平台, 以实现农业大数据的安全可靠存储、智能化管理与应用, 最终达到对农业生产过程的智能预警、智能决策和智能分析的目的, 同时为农户提供专业指导. 在以后的研究工作中, 将在 Hadoop 的农业大数据平台下对有关业务功能算法的研究, 将其中的作物病虫害检查算法、病虫害诊断算法、价值挖掘算法(CART)等算法进行分析、设计并实现并行化运行.

参考文献

1 阿尔文托夫勒.第三次浪潮.北京:新华出版社,2006.

2 农业部农业科技发展“十二五”规划(2011-2015 年). http://www.ccfc.zju.edu.cn/a/zhengcefagui/2012/0406/9978.html. [2012-04-06].

3 李秀峰,陈守合,郭雷风.大数据时代农业信息服务的技术创新.中国农业科技导报,2014,(4):10–15.

4 孙忠富,杜克明,郑飞翔,尹首一.大数据在智慧农业中研究与应用展望.中国农业科技导报,2013,(6):63–71.

5 张永军.Hadoop 分布式架构的研究与实际应用[硕士学位论文].北京:北京邮电大学,2015.

6 Fan R. Hadoop capacity scheduler. Hadoop Taiwan User Group meeting2009, Yahoo! 2009.

7 Dean J. Experiences with MapReduce, an abstraction for LargeScale computation. Proc. 15th International Conference on Parallel Architectures and Compilation Techniques. 2006.

8 Hadoop. The apache software foundation. http://Hadoop.apache.org/core.

9 周俊清.基于 Hadoop 平台的分布式任务调度算法研究[硕士学位论文].长沙:湖南大学,2012.

10 温孚江.农业大数据研究的战略意义与协同机制.高等农业教育,2013,(11):3–6.

11 Lam C. Hadoop in action. Manning Publications Company, 2010.

12 柴进.基于 Hadoop 农业数据挖掘系统的研究与实现[硕士学位论文].北京:北京工业大学,2015.

13 Bennett JML. Agricultural big data: utilisation to discover the unknown and instigate practice change. Farm Policy Journal, 2015,12(1): 43–50.

14 戴小文,漆雁斌,陈文宽.农业现代化背景下大数据分析在农业经济中的应用研究.四川师范大学学报(社会科学版),2015,(2):70–77.