



畜禽种质资源大数据平台应用场景分析与架构研究——以反刍家畜为例

张俊^{1*}, 李凯^{2*}, 邵庆勇³, 马兴跃³, 孙强³, 吴建伟¹, 王丽丽¹, 梁家充³, 李卫娟³, 欧阳依娜³, 辛苛锋⁴, 乔晓婷⁴

1. 北京市农林科学院信息技术研究中心, 北京 100097;

2. 河南省政务大数据中心, 郑州 450002;

3. 云南省畜牧兽医科学院, 昆明 650224;

4. 北京派得伟业科技发展有限公司, 北京 100097

* 联系人, E-mail: zhangj@nrcita.org.cn; leakay@hotmail.com

收稿日期: 2022-12-24; 接受日期: 2023-03-30; 网络版发表日期: 2023-06-06

国家重点研发计划(批准号: 2021YFD1200905)和云南省重大科技专项计划(批准号: 202102AE090039)资助

摘要 随着基因组学和表型组学的快速发展与应用, 研究领域产生了大量的畜禽种质资源数据, 这为构建种质资源大数据平台奠定了基础. 遗传资源保护利用与精准鉴定、种质创新与商业育种、育繁推一体化与市场监管等重要工作逐步开展, 为大数据应用提供了业务清晰且维度复杂的应用场景, 包括众多利益攸关者、巨量的多源异构数据和众多分析方法. 本研究采用领域驱动设计分析种质资源大数据应用场景, 以反刍家畜为例从数据标准规范、数据采集模式、数据模型、数据存储、数据挖掘和数据利用等多个方面开展大数据平台架构技术研究, 构建了4层结构的畜禽种质资源大数据平台, 并结合其他同行的研究成果分析讨论了下一步研究的重点与难点.

关键词 种质资源, 应用场景, 维度模型, 架构技术, 大数据平台

畜禽种质资源是畜牧业的“芯片”, 是畜牧业生产重要的物质基础之一. 《国家畜禽遗传资源品种名录(2021年版)》收录畜禽地方品种、培育品种、引入品种及配套系948个, 其中牛收录了132个, 羊收录了167个^[1]. 随着国家“种业振兴行动”和“数字乡村”的实施, 与种质资源有关的数据将会飞速增长, 如何收集、传输、存储、管理和利用这些数据, 是当前社会各界都非常关注的事情, 政府希望将数据资源变为产业发展和科技创新的动力之一, 企业希望将数据资源变为核

心竞争力甚至是数据资产, 但是如何满足这些需求还有很多工作要做. 畜禽种质资源数据可能是迄今为止最复杂的大数据类型之一, 主要包括表型组数据、基因型组数据、环境数据和管理数据等. 有组织地开展农学、生命科学、信息科学和工程科学等多学科交叉合作研究, 获取、解析、系统深入地挖掘“基因型-表型-环境型”内在关联, 逐步揭示生物重要经济学性状的形成机制, 精准鉴定识别重要性状, 对揭示作物生命科学规律, 提高作物功能基因组学和分子育种研究

引用格式: 张俊, 李凯, 邵庆勇, 等. 畜禽种质资源大数据平台应用场景分析与架构研究——以反刍家畜为例. 中国科学: 生命科学, 2023, 53: 1002–1011
Zhang J, Li K, Shao Q Y, et al. Application scenario analysis and architecture of a big data platform for livestock and poultry germplasm resources—taking ruminants as an example (in Chinese). Sci Sin Vitae, 2023, 53: 1002–1011, doi: 10.1360/SSV-2022-0279

水平等具有重大意义^[2].

本文在分析主要基因组数据和表型组数据的基础上, 根据畜禽种业的业务特点, 应用领域驱动设计, 开展场景分析归纳场景中重要的应用, 结合种质资源数据特点、国家战略、市场需求等维度综合考虑需要, 以主要反刍家畜为例, 研究讨论如何构建畜禽种质资源大数据平台(图1).

1 畜禽种质资源大数据平台现状

目前国际上集中存储DNA系列数据主要有美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)的GeneBank、欧洲分子生物学实验室(European Molecular Biology Laboratory, EMBL)的核苷酸序列存储数据库(EMBL-Bank)和日本三岛国立遗传学研究所的DNA数据库(DNA Data Bank of Japan, DDBJ), 国内主要有中国国家生物信息中心的国家基因组数据中心(National Genomics Data Center, NGDC). 这些机构和数据库都致力于通过大数据存档、管理、集成和分析提供对一套数据库资源的开放访问服务, 从而促进生命和健康科学的加速发展^[3]. NCBI维护了一组不同的35个数据库, 总共包含

36亿条记录, 其中大部分可以通过检索系统获得^[4]. EMBL-EBI拥有一套40多个开放的数据资源, 其中包含了分子生物学中的每一种数据类型. 数据资源包括沉积数据库(将实验数据存档)和增值数据库(通过提供注释、管理、再分析和集成来为归档数据增加价值), 以及开源软件工具^[5]. 基因组数据主要包括基因组DNA数据、RNA数据、蛋白质数据、代谢组数据等. 数据文件包括测序源数据、比对分析数据、注释数据、文献数据等.

表型组学源于人类疾病研究领域, 与基因组学相对应^[6], 并逐渐发展成生物学中的一个重要分支^[7,8], 在植物学领域受到了高度关注, 研究人员建立了一系列表型数据高通量采集平台^[9]. 随着畜禽表型高通量数据采集技术、设备和工程设施技术的发展, 监测和测量畜禽行为数据, 如体重、体型外貌、健康信息、生命体征信息、情绪信息、动作信息等已成为现实, 极大地推动了动物表型组学的发展^[10]. 畜禽智能化饲喂系统可以对每天监测的相关采食行为(如采食次数、采食量、采食时间等)进行全方位动态采集并实时计算采食量. 基于物联网的畜禽养殖环境数据采集能够实时获取空气温湿度、有害气体含量、可吸入颗粒物浓度、视频监控等数据. 例如, 对一个存栏量为

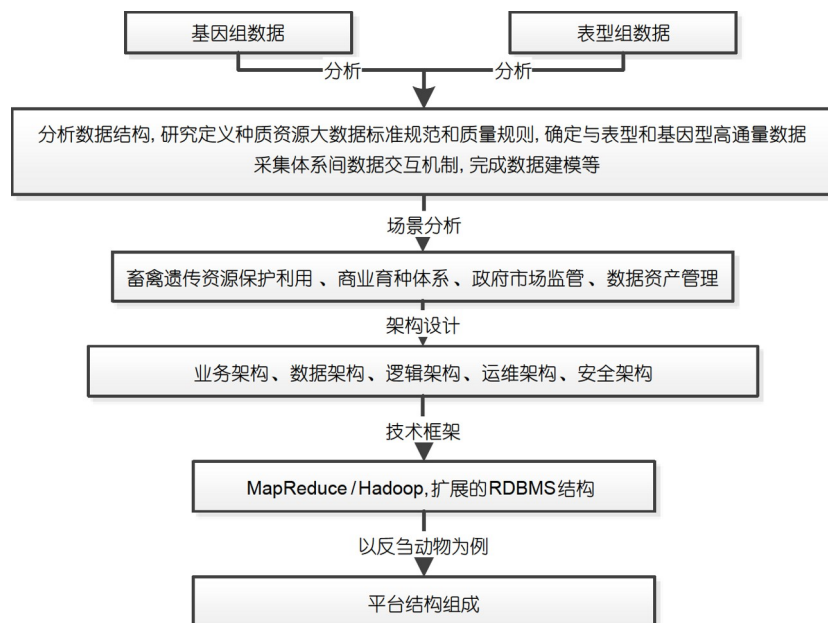


图1 畜禽种质资源大数据平台的构建

Figure 1 Construction of a big data platform for livestock and poultry germplasm resources

1000头的奶牛场进行全面的表型组学数据采集, 每天获取的数据是巨量的, 主要是非结构化的视频、音频和图片数据, 结构化的数据占比较小。根据《第三次全国畜禽遗传资源普查操作手册》的要求, 全面评价遗传资源需要调查的内容十分广泛和详实, 主要包括遗传资源概况、体型外貌(个体、群体)、体尺体重、生长发育、育肥性能、屠宰性能、胴体肌肉品质、产奶性能、长毛性能、公畜采集信息、母畜繁殖性能(个体、群体)、遗传资源影像资料等调查内容^[11]。

大数据是一种模式的转变, 涉及如何考虑数据资产、从何处获取、如何分析以及从分析中获得有价值的知识^[12]。大数据分析将面临采集问题、存储问题、管理问题、分析处理问题和领域应用问题等5方面问题, 在全生命周期的一体化协同环境下这些问题更为突出^[13]。到目前为止, 大数据分析已经在搜索排序、广告跟踪、位置与距离跟踪、因果关系发现、社会化客户关系管理、文档相似性测试、基因分析、群组发现、飞机飞行状态、卫星图像分析、CAT扫描比较、金融账户欺诈检测与干预、计算机系统黑客检测与干预、大型科学数据分析、通用名称-值对分析等方面取得很好的效果。鉴于传统的数据管理已无法满足大数据分析的需求, 有两种解决方案可供选择, 一是对原有的数据管理系统进行扩展, 使其支持对半结构化和无结构化数据的管理和分析, 二是采用新的技术模式MapReduce/Hadoop结构。大数据平台设计构建的实践没有最好模式只有最合适模式, 根据实际需要和具体应用场景选择最佳实践和应用。

2 畜禽种质资源大数据平台应用场景分析

随着“种业振兴行动”的贯彻落实, 畜禽种业发展路径清晰, 产业链正在得到前所未有的加强, 正在形成企业主导、科学家主谋、政府保驾护航的新发展格局。畜禽种业育繁推一体化体系正在快速形成, 为大数据应用提供了具体的应用场景(图2), 包括遗传资源保护利用、动物商业化育种体系、种畜禽质量市场监管、数据资产管理等。本文采用领域模型分析对以上应用场景进行分析。

(1) 遗传资源保护利用场景分析。我国正在建立以政府投入为主, 多方参与的多级动物遗传资源保护体系, 各地政府因地制宜地建立了保护区、保种场、种

公畜站和遗传物质基因库等单位开展遗传资源保护工作^[14]。河南省开展了南阳牛的遗传资源保护工作, 河南省农业农村厅建立了南阳牛保护区、保种场、种公牛站和基因库, 遗传资源规模分别为: 保护区基础母牛群体规模为3.85万头, 保种场经产母牛156头, 种牛站饲养9个血统的20头种公牛, 在基因库保存了36万份遗传材料^[15]。国家和育种企业迫切需要全面掌握畜禽遗传资源保护情况, 分析评估遗传资源保存利用效果, 实现环境信息的采集和管理, 提高保种场的生产管理效率, 降低管理成本, 并对遗传资源进行精准鉴定以期获得优异的基因或者遗传标记。

(2) 动物商业化育种体系场景分析。自我国实施“畜禽遗传改良计划”以来, 主要畜种的商业体系都得到了长足的发展, 但是与国际一流的商业育种企业相比还存在很大差距。未来我国动物商业化育种的发展目标是产业链育种和区域联合育种同时发力, 各个育种企业组建自己的育种技术服务队伍, 以活体种畜、冻精或者胚胎的销售服务为抓手, 围绕下游企业在生产管理中遇到的问题, 开展咨询、培训等服务, 推动技术服务向下或横向扩散, 形成服务链与产业链上或者区域链内育种数据采集体系重合, 形成育种大数据采集、管理、分析与利用良好互动机制, 充分利用多组学数据、表型组数据, 大力推广全基因组关联分析和全基因组选择等技术的应用, 开展基于大数据分析的育种工作(图1)。

(3) 种畜禽质量市场监管场景分析。种畜禽质量市场监管的法律体系、种畜禽商品质量规范、种畜禽产品质量检测与监管体系已基本建成。从制度机制队伍等方面不断提升市场监管能力, 从而保障了畜禽遗传资源所有者的合法权利, 同时保护消费者权益。开展种牛、冷冻精液和胚胎等遗传物质的遗传缺陷基因分子检测、亲仔关系鉴定、种公牛遗传性能检测、冷冻精液产品质量检测等, 开展有效市场质量监管, 健全奶牛遗传物质质量安全监管体系, 确保奶牛遗传物质质量安全和推广使用效果^[16]。建设集管理信息、遗传数据评估、畜禽良种登记的信息管理平台, 提高种畜禽质量水平^[17]。

(4) 数据资产管理场景分析。数据资产是拥有数据权属(勘探权、使用权、所有权)、有价值、可计量、可读取的网络空间中的数据集^[18]。充分利用畜禽种质资源大数据为其所有者创造价值, 提升数据的质量和

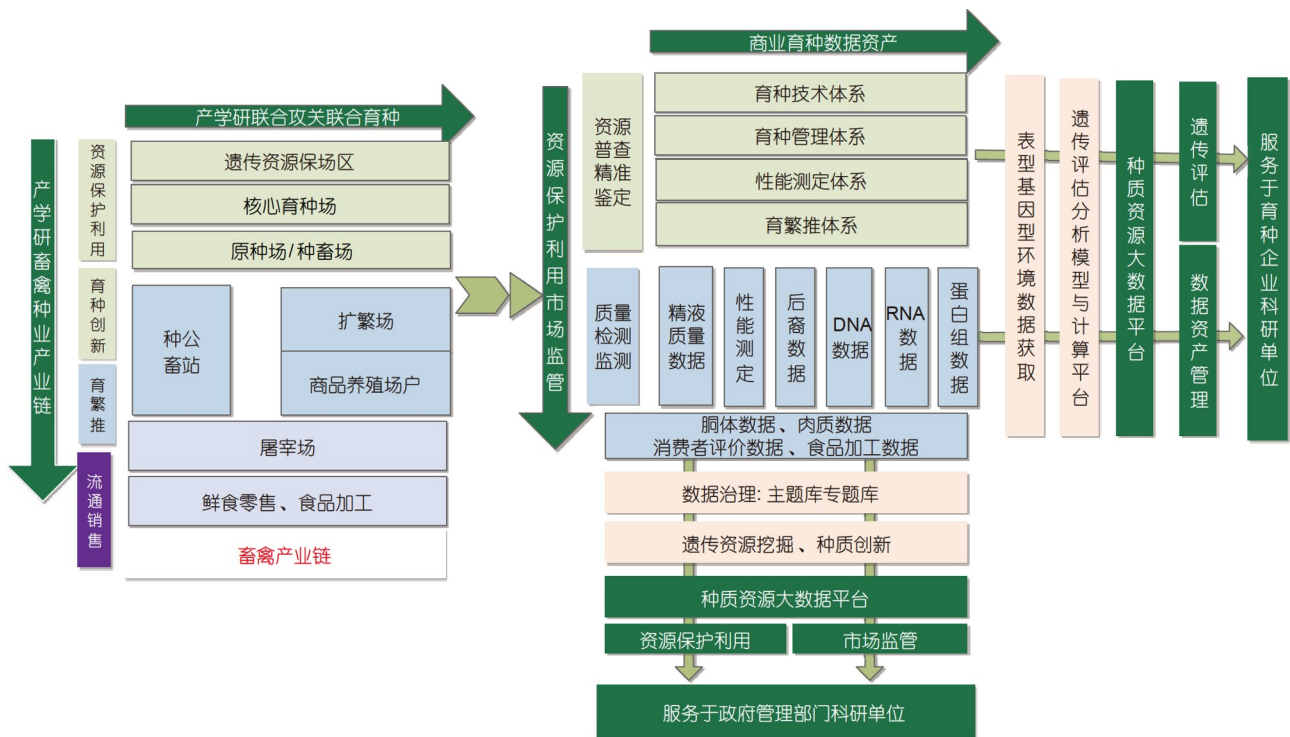


图 2 畜禽种质资源大数据应用场景图

Figure 2 Big data application scenario for livestock and poultry germplasm resources

价值,保障数据的安全.第一种方式是育种企业充分利用大数据资源与技术,低成本可持续地提升种畜禽遗传资源的性能并降低管理成本.第二种方式是直接评估数据资源价值,出售数据资源获取直接利益.在数据经济风生水起的时代谁拥有数据谁必将拥有未来的竞争优势,谁拥有具有知识产权的遗传资源和数据知识产权,就能应对新一轮国际竞争.如何在畜禽商业育种领域获得竞争优势?那就是育种企业必须具有低成本可持续地保持自己种群的生产性能优于竞争对手其他种群的能力.

(5) 用户分析. 平台的用户可以分为政府用户、企业用户和科研工作者三个类别.政府用户对于大数据平台管理的诉求主要是掌握全国畜禽遗传资源的基本情况及监测动态变化,种畜禽产品质量检测和市场监督,保障掌握核心技术和核心种源自我供给,通过政府投资促进产业高质量发展,参与国际种业市场竞争等.企业用户对于大数据平台的诉求主要是育种体系、性能测定数据、遗传评估、选种选配等常规的育种工作与数据管理,降低种畜生产管理成本、提供

种畜利用效率、加快遗传进展、提高市场竞争力与美誉度.科研工作者对于大数据平台的诉求主要是依托大数据平台提供的计算能力、数据资源和分析工具,既可作为科研平台降低科研成本,也可作为教学平台提供大数据方面教学实践.

(6) 场内或场间跨品种联合育种场景分析.随着新一轮全国畜禽遗传改良计划的实施,国家在牛羊等反刍家畜养殖企业建立了国家级核心育种场.根据育种方案,组建育种核心群,开展遗传改良工作,进行性能测定、后裔测定、遗传评估、选种选配等工作.我国反刍家畜地方品种数量众多,育种群体规模小,杂交群体规模大,遗传背景复杂.反刍家畜不同品种间主要经济性状差异很大,例如奶牛、肉牛、牦牛、肉羊、绒山羊、奶山羊,育种目标差异显著.冷冻精液、人工授精、分子标记技术和全基因组选择技术的成熟,连锁标记不平衡、QTL座位、SNP基因型等组学信息的富集,表型数据的累积,为场内或场间跨品种联合育种提供了技术可行性.充分利用大数据技术整合多源异构多场多品种多组学数据,构建联合遗传

评估方法, 开发利用我国反刍家畜遗传资源, 推进场内或者场间跨品种遗传改良工作, 将是构建我国特色遗传资源利用的有效途径, 为提高优质遗传资源的聚合进程.

3 种质资源大数据平台架构设计

架构设计的目的是降低大数据平台的复杂度. 复杂度主要来源于, 在满足遗传资源保护利用、商业化育种体系运营、政府市场监管和数据资产管理等应用场景的业务功能基础上, 平台需同时满足高性能、高可用、可扩展性、低成本、安全和规模等非业务功能.

本文以反刍家畜为例设计种质资源大数据平台(图3), 分别从数据标准规范、数据采集模式、数据建模、数据存储、数据挖掘和数据应用等层次进行设计.

(1) 反刍家畜种质资源数据描述规范. 《国家畜禽遗传资源品种目录(2021年版)》包括普通牛、瘤牛、水牛、牦牛、大额牛、绵羊、山羊等反刍家畜, 结合畜牧业、种业有关国标、行标、地方标准和《第三次

全国畜禽遗传资源普查操作手册》的有关要求, 从数据项名称(中英文)、定义、类型、格式要求、特殊说明、质量要求、取值范围等, 梳理反刍家畜种质有关数据标准规范共计有1500余项(例见表1).

(2) 数据采集模式. 不同应用场景的种质资源数据采集均可分为实时数据采集和离线数据采集(图3). 生产性能数据采集模式完全由育种企业的生产节律、育种管理体系和育种数据采集技术共同决定. 比如, 奶牛产奶性状数据的采集, 牛场的挤奶设施数量及智能化程度, 是否采用奶牛群遗传改良(dairy herd improvement, DHI)技术体系, 共同决定了数据采集模式. 例如, 山羊繁殖性状数据的采集, 主要取决于羊场查情、配种和分娩等繁殖管理模式、数据采集技术手段手动记录、APP记录或者人工智能图像识别等因素. 从以上两个例子可以看出, 种质资源大数据采集模式取决于养殖单位的现实条件, 比电商大数据、定位跟踪大数据等纯粹基于互联网数据的大数据采集模式复杂, 但是对数据更新频率要求不高, 分钟级的实时性就能满足分析需要.

(3) 数据模型. 反刍家畜种质资源大数据平台的数据类型十分复杂, 有多组学数据、表型组数据、管

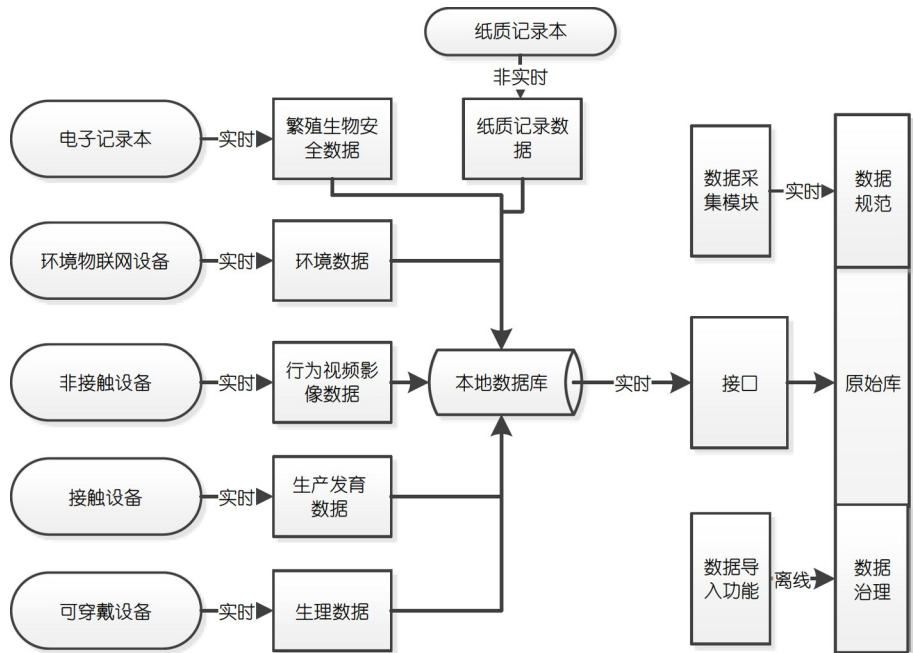


图3 数据采集模式图
Figure 3 Diagram of data acquisition

表 1 反刍家畜种质有关的数据描述规范与质量规则表

Table 1 Data description specifications and quality rules for ruminant livestock germplasm

中文名称	定义	类型	格式要求	取值范围	质量要求
个体号	平台内牛只个体唯一的标识	字符型	单位: 无; 长度15位	15位字符	15位字符
年平均温度	年平均气温, 是指全年各日的日平均气温的算术平均值. 实际上年均温度常用月平均气温来计算	数值型	单位: °C	-50~60°C	在取值范围内
母牛分娩率	分娩母牛数除以配种母牛数, 一般多用于统计一年内牛场的分娩率	数值型	单位: %; 小数点后2位	0.0~100.0%	在取值范围内
总产奶量	305 d产乳量或泌乳期产乳量(注明天数)	数值型	单位: kg; 小数点后2位	0~20万kg	在取值范围内
产毛(绒)量	所产羊毛(绒)的重量	数值型	单位: kg; 小数点后2位	0~20 kg	在取值范围内
羊毛颜色	羊毛的颜色	字符型	单位: 无; 长度4位	无	无

理数据、环境数据等. 数据模型就是数据组织和存储方法, 从反刍家畜种质资源业务、数据存取和使用角度合理存储数据. 良好的数据模型提高数据查询效率、降低存储成本、改善用户体验和减少统计分析时数据出错的概率. 本文重点采用数据仓库领域的维度模型构建数据模型, 在维度建模中, 将度量称为“事实”, 将环境描述为“维度”, 维度是用于分析事实所需要的多样环境. 比如分析繁殖性能时, 可以通过选配、选种、母畜、公畜、分娩等维度描述繁殖发生的环境, 维度所包含的表示维度的列, 称为维度属性. 维度属性是查询约束条件、分组和报表标签的基本来源, 是数据易用性的关键. 例如, 表型组繁殖数据模型(图4), 建立度量繁殖事务事实表、繁殖周期快照事实表、繁殖累计快照事实表、描述影响繁殖的维度表、用于繁殖管理的报表等.

(4) 数据存储. 在建立种质资源大数据平台时, 数据存储是最耗费时间的过程, 存储过程的开发具有挑战性, 因为不仅要考虑数据从源系统到最后的数据应用业务系统端各个存储位置和空间, 更重要的是要综合考虑业务需求、合规性、数据质量、安全性、数据集成、数据延迟、归档与世系、BI发布接口等需求, 保证在整个获取、转化、加载(extract transformation load, ETL)过程中与数据仓库保持良好的工作关系. 反刍家畜种质资源数据包括大量离线多组学数据和实时的表型组数据、环境数据和管理数据. 对于离线的多组学数据按照国际规范存储管理, 保存测序原始数据、注释数据、比对数据和分析后结果数据. 对于OLAP的存储策略将采用混合策略, 基于MapReduce/Hadoop结构, 以及传统的结构化、半结构数据

管理系统. 在适当的应用场景中统筹考虑存储高性能的相关策略, 对于结构化数据采用负载均衡、读写分离、分库分表等策略; 对非结构化数据在合适的应用场景分别采用Redis, MongoDB, HBase, Elasticsearch等策略.

(5) 数据挖掘. 反刍家畜大数据的数据挖掘方法与工具主要用于遗传资源保护利用、优异基因发现、种质资源精准鉴定、育种体系管理和政府监管等场景. 用于遗传资源保护利用领域的商业智能数据挖掘工具, 如SASS, SPSS, Clementine等. 在生物信息学领域的数据挖掘工具和分析工具, 如NCBI, EMBL和DDBJ的分析工具, 包括基因组浏览器、数据查询系统、系列比对工具等. 用于育种管理体系的数据挖掘工具主要有方差组分估计、遗传参数估计、育种值估计、基因型育种值估计, 目前都有较为成熟的商业和非商业软件MTDFREML, AIREML, DFREML, REML, DMU, VCE和ASREML等. 用于政府监管的数据挖掘工具有分类算法、聚类算法、推荐算法和深度学习等.

(6) 数据应用. 数据的价值和作用正在渗透到全社会的各行业各方面, 数据的运用、运营和挖掘利用, 正在成为一个引领社会经济发展的新动力和引擎. 有一些育种企业已将数据属性价值作为市场营销过程和售后服务中一个重要的卖点. 另外, 政府监管离不开数据.

综上, 反刍家畜种质资源大数据平台的架构见图5. 架构分为数据采集层、数据管理层、数据服务层和数据应用层.

数据采集层实现数据采集管理, 针对畜禽种业产

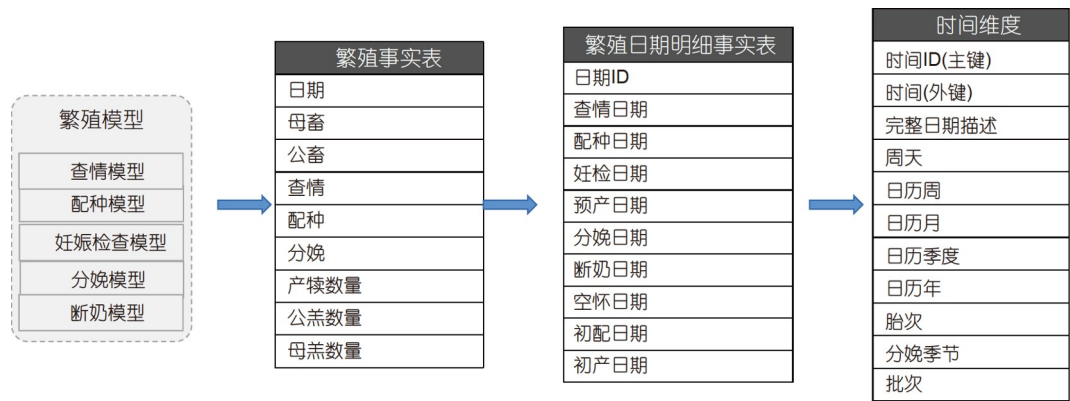


图 4 繁殖数据模型(网络版彩图)

Figure 4 Model of reproduction data (color online)

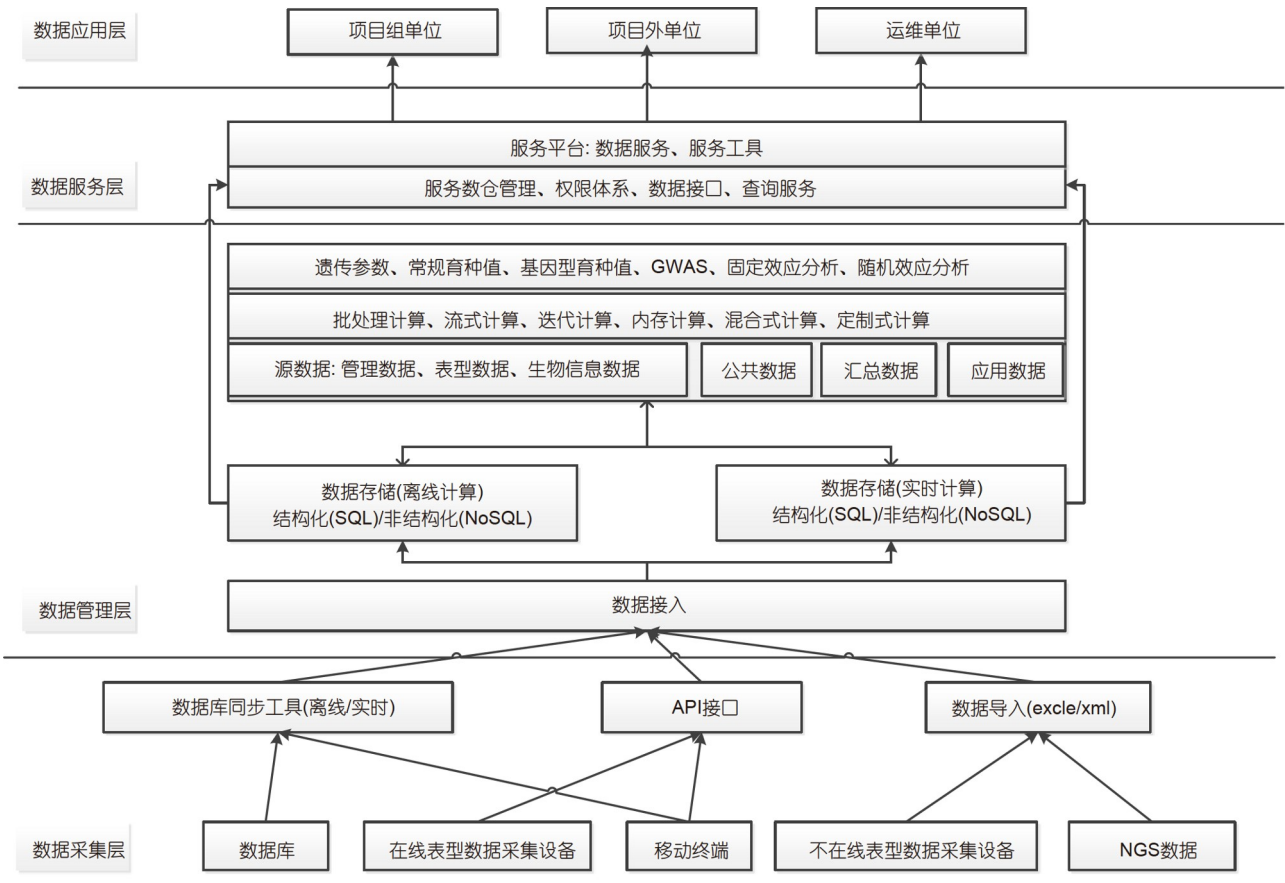


图 5 种质资源大数据平台体系架构图

Figure 5 Architecture of big data platform for germplasm resources

业链上的市场主体、科研单位和政府监管部门提供数据同步工具、API接口和数据导入功能。采用严格的数据质量控制和安全机制，保障进入平台的数据质量

和数据平台的安全性。

数据管理层实现数据存储、数据挖掘和数据分析等管理, 实现结构化、非结构化数据的存储, 支持离线

计算和实时计算, 保障ETL过程中数据一致性和准确性. 面向育种管理、科研活动和政府监管提供分析工具和计算能力, 并保证数据资产的安全性和知识产权的私有性.

数据服务层实现服务接入、服务管理、服务提供等功能, 提供数据存储、遗传评估、预测分析、资源挖掘、数据查询、服务托管等.

数据应用层面向项目组单位、项目外单位、运维单位提供相应的服务定制和服务运维.

4 讨论

产业数字化已经上升到国家战略, 数字化是未来一段时间畜禽种业发展的重要推动力, 本文以畜禽种业高质量发展的重大战略需求为研究出发点, 分析畜禽种业大数据现状和重要应用场景, 借鉴植物种质资源大数据建设的经验, 比较其他行业大数据平台技术架构, 立足于我国畜禽种业产业现状, 采用领域分析、维度建模等技术方法, 以反刍家畜为例构建了畜禽种质资源大数据平台. 国内有学者在这方面做了研究和探讨. 2020年, 马志愤^[19]在草地农业思想指导下, 以奶业为切入点, 进行草地农业大数据平台构建前期探索性研究, 利用大数据采集ETL架构、大数据处理引擎Apache Flink、大数据存储和分析等技术构建了奶业大数据平台系统, 实现奶牛场管理、牛只识别、牧场智能管理与决策和数据可视化管理, 并在197个奶牛场应用, 提高了生产管理效率, 分析了20万奶牛的关键生产性能后提出了生产性能评估体系, 为奶牛生态养殖提供了有价值的数据. 2019年, 王也^[20]设计开发了基于容器技术和流程引擎技术的家禽基因组分析流程, 实现了测序数据的预处理, 实现了系列比对、SNP变异分析、变异结构分析、RNA-seq分析、GWAS分析等功能, 并开展了家禽群体遗传分析、长

期选择对基因组影响的研究. 2022年, 马为红等人^[21]研发了商业化肉牛繁育大数据平台, 收集整理肉牛种质信息资源, 利用自主开发的智能表型设备在线自动测定肉牛表型性状, 重点支撑肉牛繁育关键过程形成肉牛种质资源大数据分析决策算法和模型, 在示范牛场应用, 为牛场的联合育种创新模式提供数据支撑. 中国家养动物种质资源数据平台系统实现了已有家养动物品种资源及种质资源的种类、空间分布、按特性类型划分的各种性状特性数据、不同品种和性别的外貌特征、特定生境下图片及视频, 以及部分种质的基因图谱等数据管理, 实现可视化查询、全文检索查询, 品种资源空间分布及属性的查询与统计分析等功能^[22]. 应用多组学海量数据分析挖掘预测绵羊的毛用、生长及繁殖性状, 根据7135只中国美利奴羊表型数据, 从基因组、转录组和表观基因组角度鉴定了与绵羊毛用和体重性状相关的遗传标记, 结合4种基因组特征(组织特异表达基因、时期特异表达分子信号、位于每个性状WssGWAS结果中解释加性遗传方差排名前100的滑动窗口, 以及合并以上三种基因组特征)构建绵羊基因组特征模型, 可以提高标准GFBLUP模型的准确性^[23].

种质资源大数据平台架构研究是平台开发的基础性工作, 为后续的技术架构、数据架构和逻辑架构的设计和平台开发奠定了基础. 畜禽种业产业链形成与演化、生物育种技术的广泛使用育繁推一体化模式的发展, 都将是平台的机会也将是巨大的挑战. 根据架构设计的“合适原则”, 立足现在预计未来, 强化标准规范的研究与应用, 跟踪了解新型表型高通量数据采集设备, 提高平台数据接口的兼容性; 加强与生物信息学、生物统计学相关工具软件的对接, 提升平台的数据分析能力和数据治理能力; 高度重平台的安全性设计, 保障用户单位的数据资产和种质资源知识产权的合法权益.

参考文献

- 1 Office of National Livestock and Poultry Genetic Resources Commission. Notice on the Publication of the National List of Livestock and Poultry Genetic Resources (2021 Edition) (in Chinese). January 13, 2021. Available from URL: https://www.moa.gov.cn/govpublic/nybzj1/202101/t20210114_6359937.htm [国家畜禽遗传资源委员会办公室. 关于公布《国家畜禽遗传资源品种名录(2021年版)》的通知. 2021年1月13日. 见网址: https://www.moa.gov.cn/govpublic/nybzj1/202101/t20210114_6359937.htm]
- 2 Zhao C J. Big data of plant phenomics and its research progress (in Chinese). J Agric Big Data, 2019, 1: 5–18 [赵春江. 植物表型组学大数据及

- 其研究进展. 农业大数据学报, 2019, 1: 5–18]
- 3 Xue Y, Bao Y, Zhang Z, et al. Database resources of the National Genomics Data Center, China National Center for Bioinformatics in 2022. *Nucleic Acids Res*, 2022, 50: D27–D38
 - 4 Sayers E W, Bolton E E, Brister J R, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 2022, 50: D20–D26
 - 5 Cantelli G, Bateman A, Brooksbank C, et al. The European Bioinformatics Institute (EMBL-EBI) in 2021. *Nucleic Acids Res*, 2022, 50: D11–D19
 - 6 Tester M, Langridge P. Breeding technologies to increase crop production in a changing world. *Science*, 2010, 327: 818–822
 - 7 Finkel E. With ‘Phenomics,’ plant scientists hope to shift breeding into overdrive. *Science*, 2009, 325: 380–381
 - 8 Bevan M W, Uauy C, Wulff B B H, et al. Genomic innovation for crop improvement. *Nature*, 2017, 543: 346–354
 - 9 Zhou J, Francois T, Tony P, et al. Genomic innovation for crop improvement (in Chinese). *J Nanjing Agric Univ*, 2018, 41: 580–588 [周济, Francois Tardieu, Tony Pridmore, 等. 植物表型组学: 发展、现状与挑战. 南京农业大学学报, 2018, 41: 580–588]
 - 10 Li Q F, Wang W T, Yu Li G, et al. Research review of information technologies for livestock (in Chinese). *China Agric Inform*, 2018, 30: 15–23 +41 [李奇峰, 王文婷, 余礼根, 等. 信息技术在畜禽养殖中的应用进展. 中国农业信息, 2018, 30: 15–23+41]
 - 11 Wang X. Operation manual for the third national survey of livestock and poultry genetic resources (first edition) (in Chinese). *China Anim Ind*, 2021, 8: 21–26 [王旭. 第三次全国畜禽遗传资源普查操作手册(第一版). 中国畜牧业, 2021, 8: 21–26]
 - 12 Ralph K, Margy R. Big data analysis. In: Wang N B, Zhou L K, Wei Z X, transl. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (Third Edition)* (in Chinese). Beijing: Tsinghua University Press, 2015. 373–384 [Ralph Kimball, Margy Ross. 大数据分析. 见: 王念滨, 周连科, 韦正现, 译. 数据仓库工具箱(第3版)——维度建模权威指南. 北京: 清华大学出版社, 2015. 373–384]
 - 13 Gong X Y, Li B H, Chai X D, et al. Survey on big data platform technology (in Chinese). *J Syst Simul*, 2014, 26: 489–496 [宫夏屹, 李伯虎, 柴旭东, 等. 大数据平台技术综述. 系统仿真学报, 2014, 26: 489–496]
 - 14 Jiang Y, Fu T, Lian H X, et al. Analysis of different ecological types of Jiaxian red cattle in different areas (in Chinese). *Acta Vet Zoo Sin*, 2014, 45: 46–55 [江燕, 傅彤, 廉礼霞, 等. 不同产区郟县红牛生态差异性分析. 畜牧兽医学报, 2014, 45: 46–55]
 - 15 Gao T Y, Bai Y Y, Sun Y, et al. The mode and application of ecological protection of animal genetic resources (in Chinese). *J Domest Anim Ecol*, 2020, 41: 8–10 [高腾云, 白跃宇, 孙宇, 等. 动物遗传资源生态保护的模式及应用. 家畜生态学报, 2020, 41: 8–10]
 - 16 Zhang S L, Sun D X. Yesterday, today and tomorrow of dairy cattle breeding (in Chinese). *China Dairy*, 2021, 6: 3–10 [张胜利, 孙东晓. 奶牛种业的昨天、今天和明天. 中国乳业, 2021, 6: 3–10]
 - 17 Xia Y Y, Lei S Y, Chen M. Thoughts on accelerating the revitalization of modern livestock and poultry breeding industry in Hubei Province (in Chinese). *Chin Livestock Poult Breed*, 2021, 17: 3–5 [夏艳阳, 雷书彦, 陈沫. 加快振兴湖北省现代畜禽种业的思考. 中国畜禽种业, 2021, 17: 3–5]
 - 18 Ye Y Z, Zhu Y Y. Data Assets (in Chinese). Beijing: Posts and Telecommunications Press, 2021 [叶雅珍, 朱扬勇. 数据资产. 北京, 人民邮电出版社, 2021]
 - 19 Ma Z F. Dairy big data platform construction and application research based on the perspective of agro-grassland (in Chinese). Dissertation for Doctoral Degree. Lanzhou: Lanzhou University, 2020 [马志愤. 基于草地农业视角下的奶业大数据平台构建与应用研究. 博士学位论文. 兰州: 兰州大学, 2020]
 - 20 Wang Y. The genetic pipelines construction and artificial selection analysis in poultry (in Chinese). Dissertation for Doctoral Degree. Chendu: Sichuan Agricultural University, 2019 [王也. 家禽遗传大数据分析流程构建及人工选择作用分析. 博士学位论文. 成都: 四川农业大学, 2019]
 - 21 Ma W H, Li J W, Wang Z, et al. Design and key technologies of big data platform for commercial beef cattle breeding (in Chinese). *Smart Agric*, 2022, 4: 99–109 [马为红, 李嘉位, Wang Zhiquan, 等. 商业化肉牛繁育大数据平台设计与关键技术. 智慧农业(中英文), 2022, 4: 99–109]
 - 22 Zhao Y G, Luo Q Y, Zheng S S, et al. Development and application of China domestic animal germplasm resources data platform system (in Chinese). *J Agric Big Data*, 2022, 4: 78–87 [赵一广, 罗清尧, 郑姗姗, 等. 中国家养动物种质资源数据平台系统开发及应用. 农业大数据学报, 2022, 4: 78–87]
 - 23 Zhao B R. Dissecting the genetic architecture of economic trait underlying fine wool sheep by using multiple omics analysis strategies (in Chinese). Dissertation for Doctoral Degree. Beijing: China Agricultural University, 2022 [赵冰茹. 利用多组学分析策略解析细毛羊经济性状遗传基础. 博士学位论文. 北京: 中国农业大学, 2022]

Application scenario analysis and architecture of a big data platform for livestock and poultry germplasm resources—taking ruminants as an example

ZHANG Jun¹, LI Kai², SHAO QingYong³, MA XingYue³, SUN Qiang³, WU JianWei¹,
WANG LiLi¹, LIANG JiaChong³, LI WeiJuan³, OUYANG YiNa³,
XIN KeFeng⁴ & QIAO XiaoTing⁴

1 Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China;

2 Henan Animal Husbandry Information Center, Zhengzhou 450002, China;

3 Yunnan Academy of Animal Husbandry and Veterinary Sciences, Kunming 650224, China;

4 Beijing PAIDE Science and Technology Development Co., Ltd, Beijing 100097, China

Owing to the rapid development and application of genomics and phenomics, a large amount of livestock and poultry germplasm resource data has been generated, laying the foundation for building a big data platform for germplasm resources. The gradual development of genetic resource protection, utilization, and accurate identification, germplasm innovation and commercial breeding, integration of breeding, propagation and promotion, and market supervision has provided a clear and complex scenario for big data applications, which includes many stakeholders, a large amount of multisource heterogeneous data, and numerous analysis methods. This research adopts a domain-driven design to analyze the big data application scenario of germplasm resources. Taking ruminant livestock as an example, the big data platform architecture technology was studied from the aspects of data standards and specifications, data collection modes, data models, data storage, data mining, and data utilization. A four-tier livestock and poultry germplasm resource big data platform was constructed, and the focus and difficulties of the upcoming research are analyzed and discussed in combination with the results from other research groups.

germplasm resources, application scenario, dimension model, architecture technology, big data platform

doi: [10.1360/SSV-2022-0279](https://doi.org/10.1360/SSV-2022-0279)