云技术下育种数据服务平台

岳 媛 赵 刚
（北京信息科技大学信息管理学院，北京 100192）

摘要：随着大数据、物联网等信息通信技术的崛起，传统育种数据管理方式已无法应对大规模育种的需求。针对现状，提出了云技术新型育种数据服务架构，并构建了育种数据服务平台。该平台以 Django 为 Web 框架，Python 为后端操作建模语言，MongoDB 为数据库，运用云存储数据中心模型及面向用户的多层服务架构为育种家提供数据存储服务和数据分析服务。平台为各育种单位开展相应服务提供了可行性的架构方案，实现了育种数据资源共享，同时推进了互联网 + 农业信息化建设的进程。

关键词：大数据；云技术；数据存储服务；数据分析服务；育种

“农以种为先”，种子是农业产业发展的首要环节和重要载体，是国内外农业产业竞争的源头和焦点。据联合国粮农组织统计，今后全球粮食总产量增长 80% 贡献率需依赖提高单产，而单产提高 60%~80% 贡献率依赖良种，因此，加快育种进程势在必行。

现代育种技术（尤其是生物技术的应用）的发展，使得作物育种数据呈现出信息爆炸的状态。育种数据不局限于单一的田间性状调查结果，同时还存在土壤、气候、水分等动态环境，影响数据、基因表达及分子标记等基因型数据，代谢物动态数据以及生产管理数据 [1]。整合和最大化利用这些生物学数据，无疑对现代育种研究具有不可估量的重要意义。

然而，调查研究发现，育种数据采集方式单一、育种数据处理手段落后、各育种单位自身条件受限，以致于无法满足育种工作的创新和新型育种活动的需要。因此，充分利用现有的信息通信技术，结合大数据、人工智能等新方法，改良育种数据管理方式，加强互联网 + 农业信息化的发展，成为首要任务。

2016 年 1 月由国家农业信息化工程技术研究中心研发的金种子育种云平台（作物育种信息管理平台）在北京上线 [2]。该平台自发布以来，有效解决了育种材料数量多、规模庞大、试验基地分布区域广等带来的工作繁重、效率不高等问题。推动我国由传统育种向商业育种、经验育种向精确育种转变，为北京建设“种业硅谷”夯实基础 [3]。

传统的育种管理平台升级为云平台，不难发现，我国育种行业的发展已经有所进步，但在更深层面上，育种行业仍然只是行业而并未形成产业，与世界的差距依然存在。对比我国顶尖种业公司登海种业及跨国种业孟山都，分析二者经营规模的差异得知，孟山都种业销售收入总趋势是逐年递增，而我国登海种业以及大部分种业的销售收入情况增长仍然不稳定 [4]。造成这种差距的主要原因是我国种业员工文化程度低、品种审定制度门槛过高、海量数据处理较慢、缺乏统一的数据分析平台等。可见，国内育种缺少的不只是强大的育种技术，更是一种解决传统问题的创新思想。

基于此，本文密切迎合育种行业需求，结合物联网、大数据技术、人工智能及机器学习方法，提出并构建基于云技术新型架构的育种数据服务平台，研究新型、高效的育种数据管理和数据分析方法，一方面可以提高育种工作人员的效率，研究出更加优质的作物品种；另一方面主动革新育种手段可以提高育种企业的竞争力，打响国内种业品牌。此外，将人工智能领域的机器学习算法应用在育种数据管理中，响应了国家所倡导的“三农政策”，将互联网与传统农业深入结合，缔造出新的农业发展

基金项目：北京市科委重大项目（D151100004215003）；国家自然科学基金资助项目（61572079）；北京市社会科学基金研究基地重点项目（17JDGLA037）
通信作者：赵刚

态势。

1 构建基于云技术新型架构的育种数据服务平台

本文结合云技术进行具体研讨，将新型架构部署至私有云中，从而实现依照相应的付费标准为更多的企业提供服务，实现育种资源共享。图1为云技术下育种数据服务平台总框图。在此过程中，数据的存储及分析服务成为设计焦点。

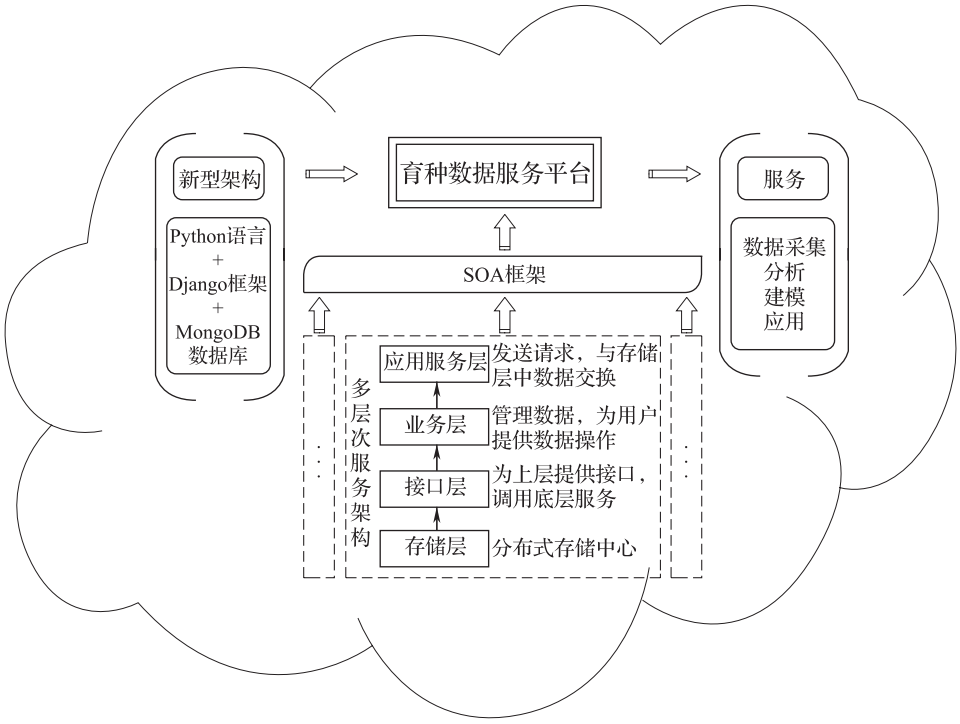

图1 云技术下育种数据服务平台

1.1 平台需求分析概况 育种数据服务平台的用户主要为育种工作人员、育种科研机构和平台管理人员等，提供的主要服务是对育种数据进行管理，涉及到育种数据的采集、数据分析和模型应用等一系列过程。用户在获得平台登录许可后，可以根据需求对其操作，如获取实时育种性状数据、天气以及地理属性数据；根据需求对数据进行图形化展示，方便用户重点分析数据潜在规律；平台采用机器学习算法和大数据技术，对数据进行客观分析，以便为用户提供合理的决策意见。此外，该平台搭建在云技术上，方便不同区域的用户能够随时获取育种数据，还可更好地实现育种数据的共享，为育种科研工作提供支持。

随着现有育种数据的增长，目前育种软件不能很好地处理这些数据，因此需要该平台的新型架构作为底层框架，支撑数据分析及建模的运行。

1.2 新型架构的组成及优势 美国孟山都公司采用传统的SQL Server+IIS+EX为用户提供服务。本文提出的架构以Django为Web框架，Python为后端操作建模语言，MongoDB为数据库，从而提供快速存储服务及数据分析服务。

1.2.1 Django框架 为更好地提供育种数据分析服务，选取机器学习算法，Python语言类库多、调用方便的优点更能适用于育种数据服务平台的开发。Django作为Python的一个开源框架，应用开发遵循MVC模式。其中C是应用程序中处理用户交互的部分，而Django更关注的是模型（Model）、模板（Template）和视图（Views），因此Django也被称为MTV框架。该框架分工明确，层次分明，代码相互不交叉，只需解决连接问题。同时基于Django框架的育种部署系统具有实用、开源、轻量级等多个优点，可方便地移植于Windows、Linux等多种操作系统平台，在云平台上充分发挥作用，为育种专家提供有效的数据分析服务。

1.2.2 MongoDB数据库 近年来伴随高通量测序技术的发展，产生了海量作物育种相关基因及其

表达数据，形成了育种大数据。为满足数据的存储效率及育种数据服务的相关要求，研究了 MongoDB 数据库与育种数据之间的相关性，运用其面向文档的数据存储模式和可扩展的表结构，实现提高育种数据读取和存储的速度，为育种数据分析提供操作便利性和可解读的数据存储结构，并且大大降低数据的复杂度和关联度，使其更加适用于育种。与孟山都公司采用传统的数据库相比，本架构充分利用 MongoDB 的 NoSQL 数据库特性，在安全上有效防范传统 SQL 注入，解决相关育种数据平台的数据安全问题。

同时，把 MongoDB 部署在云端，企业可以在世界范围内存储更多的数据，吸引更多的育种机构，关联更多的信息，创造更高的价值。在云技术环境下的 MongoDB 发挥其原生的可扩展框架，保持育种数据的可用性和完整性的自动管理，还有可启用的分片和水平扩展技术，提供了云存储所需的技术。设计实现中，利用 MongoDB 对 MapReduce 的支持及其 Hadoop 接口，设计便于开发及扩展的育种数据服务平台。

1.2.3 云存储 对于育种数据的分析，单单几次用户操作轨迹的跟踪并不能准确地推算出用户的行为习惯，几天的系统日志分析结果并不能让观察人员做出最合理的决策，只有通过大样本随机对照双盲测试才能断定某种商品的价值[5]。因此对于采集到的海量数据，MongoDB 成了存储的最佳选择。此外，为保证全国地域的数据采集及数据的时效性，还需构建一个快捷且稳定的网络数据集的存储基地——云端存储中心。无论何时何地，研究者都可以通过云服务把最新数据存储在云端，也可以获取其他地域的最新数据，MongoDB 没有给出存储上限，随着数据采集周期性的加长可得到更多数据，在使用诸如育种决策系统等分析系统对大数据进行处理时，能带来可靠的分析结果，便于做出正确的决策。

1.3 系统设计逻辑 育种数据服务平台采用 Django 框架和 MongoDB 相结合的新型 Web 架构。既确保各功能模块之间互不影响又提高了育种数据读取和存储的速度。在此架构之上，平台提供从数据采集到数据分析、数据处理等一系列功能。其中数据存储服务使用新型流数据技术并由云技术的 MongoDB 提供，从而解决存储速度慢、容量小等问题。

1.3.1 系统功能模块 平台将搭建于云服务器上，利用云存储的可扩展性和高访问特性，实现育种数据的海量管理和共享机制。其育种平台总体结构图如图 2 所示。

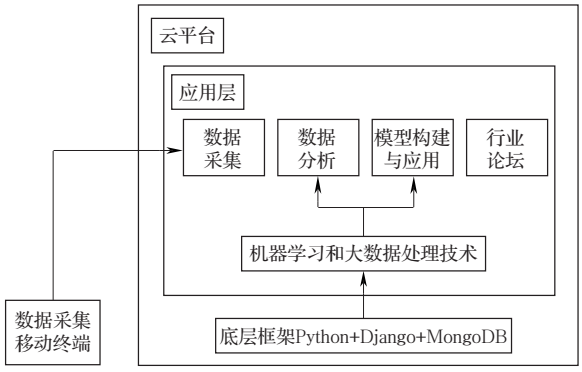

图 2 育种平台总体结构图

数据采集过程，分为在线数据分析和离线数据分析；在线数据为实时接收移动终端传送的数据；离线数据包括 .xls 和 .txt 格式的数据。

数据分析过程，主要采用机器学习和大数据技术对获取的数据进行分析；机器学习算法主要包括支持向量机、朴素贝叶斯的分类算法和 K-means 聚类算法等；大数据技术采用的是 Spark 技术对数据进行运算。

模型构建与应用过程，主要是通过机器学习算法分析数据后，构建合理的数据分析模型。用户可以通过构建的模型对新的数据进行分类或者预测等。

行业论坛，这一模块主要为育种人员提供交流的窗口，分享育种经验。

1.3.2 系统核心技术及实现 利用 Python 提供的 pickle 类将代码中建立好的模型，从字节流转成字符流文件，并将其存在文件系统中。其优势在于无需重复建模，只需调用即可。

1.4 运行效果 数据统计、数据分析、结果展示分别如图 3、图 4、图 5 所示。

2 面向企业的育种数据服务

2.1 面向用户的多层次服务架构 云存储是面向用户，以服务为中心的存储管理，其特点为按需服务，自动化运维。本文设计的非结构化数据的云存储架构建立在 Hadoop 之上。层次结构主要包括以下部分。

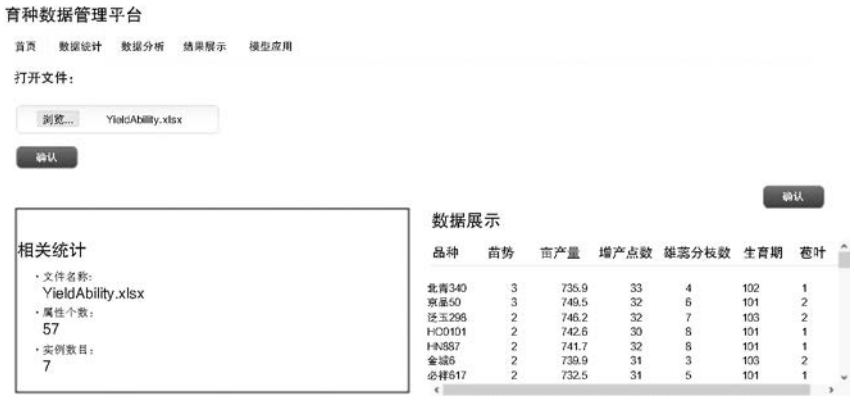

图3 数据统计

图4 数据分析

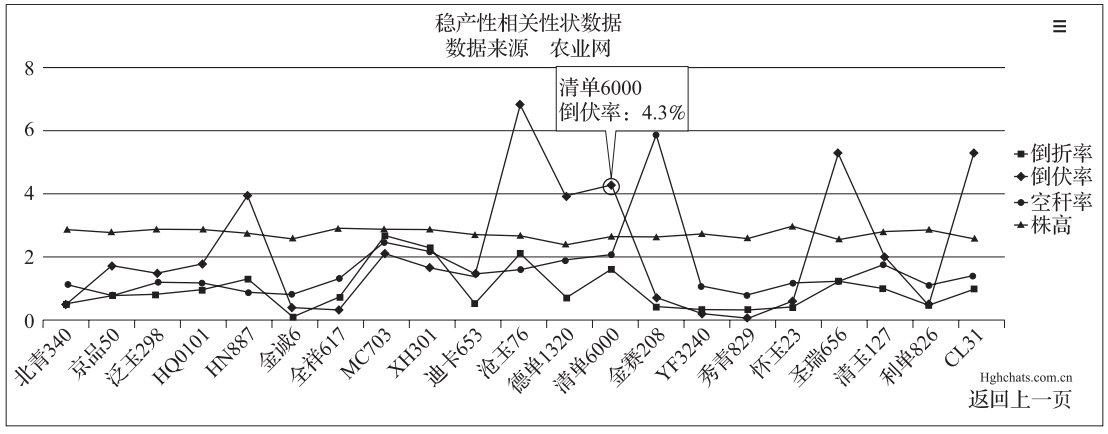

图5 结果展示

存储层：位于多层次服务结构的最底层，起存储数据作用。对于育种企业，产生的育种数据越来越多且用户量较大时，传统的单节点存储已不再满足需求，采用多节点存储方式对数据进行有效的管理，搭建分布式数据存储中心，将数据分散存储的同时对外提供了更专项化的服务。

对于存储层的数据管理，各育种企业通过物联网技术将采集的育种数据上传至分布式数据存储中心，育种专家或系统管理人员将相关数据进行收集，后台数据库 MongoDB 处理大量的流数据，也为大数据的分析提供了强有力的保障，并提供了 Hadoop 接口，能与第三方数据分析工具完美结合。

接口层：为上层（业务层）提供接口，调用最底层（存储层）的数据和方法。其中数据采集接口服务[1]主要包括性状数据采集设备、田间视频监控设备、生长环境信息采集设备。用于数据分析接口服务包括自动化考种系统、育种试验分析软件以及支付系统等。

业务层：即育种系统业务逻辑设计，为用户提供数据操作，并完成用户的请求。

应用服务层：用户与云存储数据中心的集群进行交互，发送相关请求，并与存储层中的数据交换实现数据操作。

2.2 云存储数据中心模型 云存储是云计算的延伸，它致力于解决云计算中海量数据存储的问题。通过互联网的连接，云存储为用户提供了访问共享存储池的能力。用户可随时随地进入云平台，享用该服务。面对采集到的农作物数据，分散在各个科研单位的数据集，为了更好地融合这些数据集，要求数据存储系统的设计需满足：可扩展性，支持海量数据处理，实现资源的按需扩展；可靠与可恢复性，在进行种子培养、性状采集时会产生相应的数据，部分原始数据甚至还具有不可重现性，这要求存储系统必须具备较强的可恢复性，能够实现数据灾备和恢复；高访问性，伴随育种协作的不断推进，不同科研单位对于数据的交互需求逐渐增加，数据量也随之增长，这需要系统具有较高的访问性能，能够在很短的时间内传输并反馈海量的数据。

云存储数据中心是由多个物理机组成的集群系统[6]，如图6所示，提供可扩展、高可靠性及高访问性存储空间。

该集群类似于亚马逊简单存储服务（Amazon S3），由一个控制节点和多个数据节点构成，对育种用户提供统一的管理和维护。其中，控制节点用来保存和管理种子性状的元数据信息；而数据节点则用来存放真实的数据，如PDF文件、Word文档、视频文件等。数据存储中心向育种用户提供统一的服务接口，用户通过标准化服务接口操作育种信息的存储、删除、移动、计算处理等任务。该集群系统存在于基础设施层，具有可靠性和鲁棒性，各个节点相互独立，一个节点的损坏不会影响其他节点。当控制节点出现问题时，整个系统会停住工作，此时系统中的备份控制节点发挥作用，迅速完成数据恢复。待恢复完毕，系统继续服务。

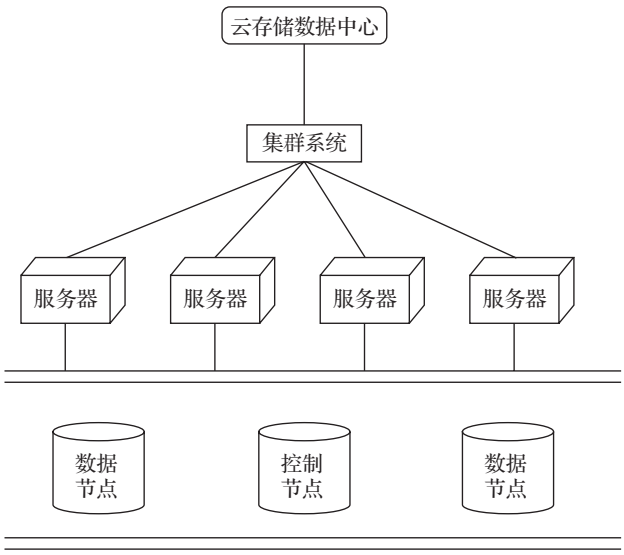

图6 云存储数据中心模型

3 云技术平台的应用

3.1 育种类数据需要“云架构”的支撑 综上，云技术除解决了育种类数据“大规模、非结构化、数据处理慢”这3个传统特点外，还有数据分析中算法需要“云架构”的支撑[7]。云计算具有低成本、易扩展、计算能力强等优势，将育种数据上传后选择云中相应算法得出性状对应的结论。对比国内先锋“金种子云平台”提供的种质资源管理、试验规划、性状采集APP、品种选育、品种区试、系谱管理、数据分析、基于电子标签的育种全程可追溯等服务。其中，本平台数据分析模块采用机器学习和大数据技术，包括支持向量机、朴素贝叶斯的分类算法和聚类算法等；同时也采用了Spark技术对数据进行运算。方法诸多，便于使用。

3.2 除育种行业外其他行业的迁移 该平台存在于云上，主要为育种行业提供服务[8]，相关数据存储、数据分析、模型应用等模块加强传统育种信息化建设，向互联网+农业迈进。当然，其他行业的云平台也可迁移至此。例如零售业，商家对卖场物品的摆放可运用数据分析手段对商品间进行相关性分析；投资方对卖场的选址可用聚类算法；利用NLP挖掘潜在客户，从而刺激销量；运输行业的运输路线等。可见，该平台亦推进了其他行业的发展进程。

苜蓿种质资源概况及耐盐性研究进展

陈小芳 徐化凌 毕云霞
（山东省东营市农业科学研究院，东营 257091）

摘要：苜蓿是世界上种植最广的优质豆科牧草之一，培育耐盐新品种是减少盐渍化土壤对苜蓿生长发育和产量影响的重要途径。综述了苜蓿的起源、分布和分类，耐盐生理及耐盐性育种方面的主要研究进展，并对其耐盐性研究的前景进行了讨论。

关键词：苜蓿；种质资源；耐盐生理；耐盐育种

盐渍化土壤使农作物低产或不能生长，是制约现代农业增产、增效和实现农业良性发展的主要限制因素之一，是经济与社会问题，同时也是生态环境问题。通过工程措施解决盐渍化是一项复杂、难度大、需时长的工作，因此培育耐盐品种就成为改良盐渍化土壤的有效途径。通过对苜蓿耐盐性的提升，不仅可使饲草产量增加，使蛋白质饲料短缺问题得以缓解，且可使盐碱地具备更高的利用率，对生态环境保护、农业和社会经济可持续发展有着重要意义。

苜蓿是栽培历史最为悠久的豆科牧草，在全球范围内均有种植，可较好的生长于中性或轻度盐碱地上，属于相对耐盐的一类豆科牧草。苜蓿茎叶繁茂，地表覆盖度大，可有效减少地面蒸腾，抑制土壤返盐，同时其发达的根系入土深，可促进降水的淋盐作用，使土壤盐分含量下降，并减少水土流失，具有广泛的农学意义。

基金项目：东营市科技计划专项（2015GG0103）；山东省农业良种工程项目（2016LZGC010-5）经费资助

1 苜蓿的起源、分布与分类

1.1 苜蓿的起源及传播 苜蓿是一种来自近东和

4 总结与展望

因现有育种数据管理工具的落后，育种数据分析平台的匮乏，提出并构建了以 Django 为 Web 框架、Python 为后端操作建模语言、MongoDB 为数据库的新型架构，论述了其在育种平台下的应用优势。随后，研究该架构下的育种服务与云技术模式的契合点，在云技术下重点开发了数据统计、分析、模型应用等服务，育种家可根据种子特征因素绘制种子性状统计图，进行数据分析，还可对种子材料的优劣进行评价并利用机器学习算法对其性状聚类，挖掘潜在规律。此外，还重点建设云存储数据中心及对外分析服务，解决了海量数据存储的问题，其特色在于便利了育种数据的管理，减少了育种数据存储的成本，实现了育种数据的资源共享。具有广阔的应用前景。

目前，网络安全漏洞、数据泄露、存储故障等问题日益加重，安全性成为制约云存储发展的首要问题，如何改善安全性也迫在眉睫。云存储中数据的安全性尚待进一步研究。

参考文献

[1] 刘忠强，王开义，赵向宇，等. 云环境下作物育种信息化模型研究[J]. 农机化研究，2017（3）：7-11，21
[2] 张卫，韩义雷，操秀英. 金种子育种云平台发布助力育种管理工作升级为云服务 [J]. 蔬菜，2016（2）：12
[3] 赵广飞. 北京建成全国首个大型育种云服务平台 [EB/OL].（2016-01-19）[2018-07-12]. http：//www. xinhuanet. com/politics/2016-01/19/c_128642244. htm
[4] 王彤. 我国种子行业现状对比分析及改进建议 [J]. 商场现代化，2016（20）：246-247
[5] 何倩，陈亦婷，董庆贺，等. 基于 MongoDB 的物联网接入云服务平台 [J]. 桂林电子科技大学学报，2017，37（1）：1-7
[6] 胡珊珊. 面向云存储的非结构化数据存储研究与应用 [D]. 广州：广东工业大学，2014
[7] Mao X，Zhao G，Sun R Y. Parameter optimization based two-layer SVM classification model for evaluation of maize breeding[C]. 13th International Conference on Natural Computation，Fuzzy Systems and Knowledge Discovery，2017：2497-2502
[8] Roopaei M，Rad P，Choo K K R. Cloud of things in smart agriculture：intelligent irrigation monitoring by thermal imaging[J]. IEEE Cloud Computing，2017，4（1）：10-15

（收稿日期：2018-07-12）