专题：大数据与云存储

Big Data and Cloud Storage

客座编辑

张广艳（1976– ），清华大学计算机系长聘副教授，国家杰出青年科学基金获得者。目前担任CCF信息存储技术专业委员会副主任、CCF计算机历史工作委员会副主任、CCF系统软件专业委员会委员。主要从事大规模数据存储与分析的理论和方法研究，包括存储系统、数据压缩、大数据计算、AI计算系统等方面。研究得到包括国家杰出青年科学基金、国家重点研发计划、中国工程院战略研究与咨询项目、973和863等国家科研项目的支持。发表学术论文60余篇，其中包括FAST、SOSP、USENIX ATC、EuroSys、*ACM TOS*、*IEEE TC*、*IEEE TPDS*等计算机系统领域顶级国际会议和期刊论文20余篇。以第一发明人获得美国发明专利授权、中国发明专利授权10余项。研究成果应用到多家国内骨干企业的存储产品和生产系统中，效果良好。

导读

随着信息化浪潮的迅猛发展和人工智能技术的快速迭代，大数据与云计算已成为支撑前沿信息技术的新型基础设施。大数据蕴含着巨大的信息价值，被誉为数字经济时代的“石油”，成为推动经济社会发展的关键生产要素之一。云计算系统能够为大数据提供全周期、低成本、可扩展的存储、处理和服务，成为大数据价值发现与实现的重要平台。云存储作为云计算系统的重要组成部分，基于海量、异构的硬件构建分布式存储，为云计算系统提供高性能、高可靠和可扩展的数据存储服务，有效支撑大数据的可靠存储与高效分析。大数据与云存储可以为国家重大战略应用和人民群众生产生活提供服务，将成为未来数字时代新的发展引擎和动力源泉。国内外企业巨头纷纷着力布局大数据与云存储相关产业，我国政府也大力支持大数据与云存储产业的发展。

大数据与云存储领域涵盖硬件、系统、软件、算法和应用等多个方面，对计算机信息系统全栈提出了一系列挑战。为了针对性地应对大数据与云存储发展中的技术挑战，本刊特组织“大数据与云存储”专题，探讨学术界和产业界关注的技术问题和解决方案，旨在加强对大数据与云存储产学研发展的理解与认识。经过同行专家评审，最终录用了5篇文章，主题涵盖行业大数据的存储和访问、分布式系统中的低效查询优化、时序数据的高效压缩、分布式计算环境中的任务与资源匹配、碳排放量预测等方面。

针对湍流大数据的存储、访问和管理存在数据集成困难、数据访问低效和存储效率低的问题，张晓等人设计了一个面向湍流大数据的分布式文件系统TDFS。为了解决大规模分布式系统中存在低效行为的问题，杨海龙等人提出了一种通用的低效查询语句检测和优化流程，并总结了4类显著影响大数据应用性能的低效行为和对应的优化策略。为了解决现有压缩策略缺乏针对NVM与IoT时序数据特征的优化机制的问题，蔡涛等人提出了一种面向NVM的IoT时序数据多态协作压缩策略。为了解决广域分布式计算环境中任务、资源匹配的响应延迟高、匹配效率低的问题，肖利民等人提出了一种面向广域分布式计算环境的任务与资源动态匹配方法。为了实现对钢铁产业碳排放量的预测，李凤云等人提出了一种基于长短期记忆网络的炼钢厂碳排放量预测方法。

大数据与云存储潜在经济价值和社会价值的实现程度，取决于基础设施建设的完善程度及核心技术突破的进展速度。目前，大数据与云存储行业已经实现了一定规模的部署和应用。然而，作为一个庞大的生态系统，要实现高质量的全面发展，仍需克服许多必要条件和核心技术的挑战，需要学术界和产业界进行更多的技术协作和攻坚。本专题以“大数据与云存储”为主题，围绕大数据与云存储中的行业大数据存储、分布式系统查询、数据高效压缩、任务与资源匹配、碳排放量预测等进行讨论。希望通过这些讨论，抛砖引玉，启发更多政策思考、关键技术进步，赋能大数据与云存储产业的持续健康发展。