

省部级重点实验室

数据工程与知识工程教育部重点实验室

一、实验室简介

“数据工程与知识工程教育部重点实验室”是教育部批准成立的实体性科学研究机构，以中国人民大学信息学院和信息资源管理学院为依托建设单位，是国家科技创新体系的重要组成部分，也是国家在数据工程与知识工程领域，组织高水平基础研究和应用基础研究、聚集和培养优秀科学家、开展学术交流的重要基地。

1. 历史沿革

2009年2月19日，教育部印发《关于“高可信软件技术”等教育部重点实验室通过验收的通知》，中国人民大学第一个省部级重点实验室——数据工程与知识工程教育部重点实验室，于2008年10月27日经教育部专家组验收一致通过，正式挂牌运行。数据工程与知识工程教育部重点实验室开创了人民大学发展史上的又一个“第一”，也在新中国科学技术发展史上写下了具有特殊重要意义的一笔。在此之前，省部级以上重点实验室100%是自然科学类的，人民大学数据工程与知识工程教育部重点实验室的建立，标志着以“文理结合”为特点的重点实验室实现了“零”的突破。

2. 实验室主任

数据工程与知识工程教育部重点实验室的主任由杜小勇教授担任，杜小勇教授现任中国人民大学校长助理、杰出学者特聘教授，是国务院特殊津贴专家，担任中国计算机学会大数据专家委员会主任、工信部国家大数据标准工作组副组长兼大数据治理研究组组长、《大数据》副主编、ACM Transactions on Data Science 编委等。杜教授在杭州大学、中国人民大学和日本国立名古屋工业大学获得本科、硕士、博士学位并分别担任过助教、讲师、助理教授等职务，1999年起任中国人民大学教授。杜小勇教授的主要研究领域是数据库与大数据，其团队研制的国产数据库 Kingbase



ES 系统先后获得多个省部级和国家科技进步奖。发表学术论文 300 余篇，获国家发明专利 10 余项，承担了多个国家级科学项目。

杜小勇教授曾担任十五 863 计划先进制造领域“数据库管理系统与应用”重大专项专家组组长，863 计划软件重大专项专家组成员。曾担任 863 计划项目、国家重点研发计划项目、核高基课题、973 计划课题等负责人。现担任科技部“云计算与大数据”重点研发专项总体专家组成员，教育部科技委信息学部委员。担任多个国内外学术期刊编委，数十次担任重要国际学术会议程序委员会主席或委员。

3. 环境及设备

实验室具有相对独立的人事权和财务权，为独立的预算单位，在资源分配计划上单列，与中国人民大学各学院平行。实验室在各类科技计划以及“985”工程和“211”工程经费的支持下，已经初步建立了一个可提供标准化 IaaS 服务的云计算平台——人大行云。目前，云平台的服务器节点数达到了 106 个（其中刀片式服务器 32 台，2U 机架式服务器 74 台）；聚合计算能力达到了 13Tflops；存储能力达到了 770TB。

4. 成果概述

实验室以数据工程与知识工程领域的理论研究、工程开发和人才培养为核心，集中资源，在数据工程与知识工程的重要领域实现重点突破。截至 2020 年底，实验室先后承担了国家重大重点科研项目（核高基项目、863 项目、973 项目、国家重点研发计划项目）40 项，国家自然科学基金项目 150 项，国家社会科学基金项目 33 项，承担各类科研项目总数达到 954 项，项目总经费达到 4.2 亿元。同时实验室科研人员累计发表被 SCI、EI 收录的论文共 1298 篇次；共获授权专利 95 项；获授权软件著作权 49 项；出版各类教材、专著 94 部。实验室取得的研究成果先后获得国家科学技术进步二等奖、教育部科学技术进步一等奖、二等奖，北京市科学技术进步二等奖（两次），中国计算机学会科学技术进步一等奖，有效缩短了我国在数据工程与知识工程领域同国际水平间的差距。

二、代表性成果与案例

1. 国产数据库 KingbaseES 研制

(1) 成果描述：

实验室与人大金仓公司开展产学研密切协作，在国产数据库管理系统内核研制、XML 数据和关系数据的统一管理、海量数据的联机分析加速等方面取得了一系列创新性研究成果。实验室的相关研究成果突破了数据库管理系统“三高一大”（高可靠、高性能、高安全、大数据）核心技术，并研制了具有自主知识产权、安全可靠、自主可控的数据库管理系统 KingbaseES。



中国人民大学—腾讯协同创新实验室成立于 2020 年，实验室主任为杜小勇教授

数据库是高度复杂的软件系统，是信息系统的底层支撑技术，是国家实施数字化转型的基础软件之一。然而，数据库相关核心知识产权却长期被国外厂商所垄断，造成了严重的不可控风险且经济开销庞大，数据库技术已成为制约我国数字化发展的“卡脖子”技术。中国人民大学与腾讯公司数据库团队紧密合作，聚焦金融级分布式数据库（TDSQL），在系统强一致与高可扩展、HTAP 系统、新硬件加速数据库等方面开展创新性研究，中国人民大学数据库团队研制的“支持多级一致性的高可扩展分布式数据管理”、“基于全时态数据模型的高性能融合引擎”等技术，



已部分落地到 TDSQL 数据库中，提升了 TDSQL 的技术水平，为 TDSQL 顺利通过工信部组织的“信创分布式数据库系统测试”和支撑第 7 次人口普查提供了重要的技术支撑。双方自合作以来，已完成或正在执行 10 项重要项目，协议科研经费总额超 600 万元，提交或获得授权专利 10 余项，发表或录用多篇 CCF A 类论文。

时序数据库是面向 IoT 和 AIops 应用的一种新型数据库产品，近 3-5 年来，国内涌现出来了 TDEngine, IoTDB, DolphinDB, MatrixDB 等时序数据库产品，互联网大厂华为、腾讯等也有相应的时序数据库产品。随着智能制造、云计算等应用的不断深入发展，这类数据库成为近期风险投资的特点。然而如何有效评估这类数据库产品的性能，指导时序数据库相关技术和产品健康发展，一直缺少有效的评测基准。实验室从 2018 年开始研发的时序数据库基准 TS-Benchmark 最近在 CCF A 类会议发表，也引起国内外学术界和厂商的关注。目前，行业内较为领先的产品华为时序数据库、TDEngine、IoTDB 都应用了 TS-Benchmark，并参与到该基准的改进中。TDEngine 已经开始走向国际市场，未来几年我国在时序数据库领域有望打造出一批国际领先的产品。

(2) 应用获奖：

KingbaseES 数据库管理系统在十多个行业领域和六十多个全国性重大信息化工程核心关键业务中得到了规模化应用，推广应用 50 余万套，在国产数据库信创市场占有率超过 50%，研究成果“数据库管理系统核心技术的创新与金仓数据库产业化”获得 2018 年国家科技进步二等奖。

(3) 成果案例：

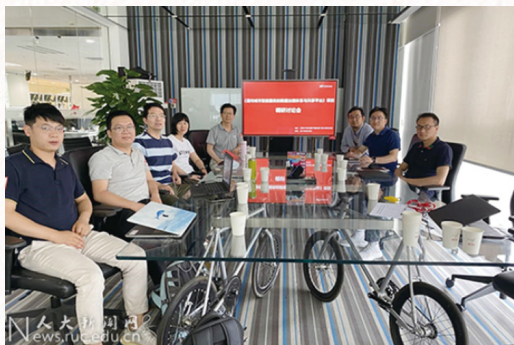
实验室依托人大金仓公司积极开展成果转化，专注数据库产品开发与服务。金仓数据库 KingbaseES 已经成为应用最广泛的国产数据库知名品牌，全国累计部署超过 50 万套，遍布全国 3000 多个县市。广泛应用于电力、司法、军工、金融、电信、教育、审计、国土、信访等超过 20 个重点行业，并被中组部、国家电网、预警机等国家要害部门采用。目前该企业估值超 4 亿元，成为国产数据库的标志性企业之一。

2. 数据治理关键技术研究

(1) 成果描述：

实验室关注十九届四中全会提出的“国家治理体系与治理能力现代化”要求，从政府数字化转型这一时代重大命题出发，着眼于以“城市大脑”为代表的智慧城市建设，结合智慧城市建设的前期实践，总结凝练出城市数据治理方法论体系并提供治理工具。具体而言，实验室对跨区域跨行业政务服务协同所面临的多领域异构系统对接、互通互联与安全可信等问题，构建了基于区块链的、支持跨城市政务服务应用的可信数据共享平台，实现基于智能感知的综合性城市精细化管理服务，并进行应用示范验证。

此外，实验室积极参与物联网和智慧城市及社区的数据处理与管理国际标准制定，中国人民大学信息资源管理学院安小米教授受 ITU-T FG-DPM 主席邀请自 2017 年 7 月开始参加 ITU-T FG-DPM 工作，参与国际电联电信标准化局 (ITU-T) 物联网和智慧城市及社区 (SC&C) 数据处理与管理焦点组工作。团队共完成了 4 个国际标准，占整个中国团队的 50%，例如，ISO 37156:2020 Smart community infrastructures — Guidelines on data exchange and sharing for smart community infrastructures；同时，研究团队牵头制定了国内标准，如“智慧城市基础设施 - 突发公共卫生事件数据高效利用指南（2020）”。此外，实验室牵头制订了一项国家标准，即 GB/T 34950-2017 非结构化数据管理系统参考模型，参与编制的其他国家标准 13 项。





在技术方面，数据融合是高质量数据分析的基础与前提，是数据治理的重要内容。其目标是整合多源异质数据，形成统一的数据视图。人在回路的数据融合技术近年来备受学术界与工业界的重视。图灵奖获得者 Michael Stonebraker 教授提出构建第三代人在回路的数据融合系统。在工业界，包括沃尔玛、阿里巴巴在内的多家企业也尝试利用人在回路的方法解决大规模的数据融合难题。重点实验室近年来围绕数据融合方向做了一系列原创新研究，相关成果共发表中国计算机学会推荐 A 类论文 40 余篇，在 SIGMOD 2017、KDD 2018 和 ICDE 2019 多个 A 类会议上组织以人在回路数据融合为主题的辅导报告，并成功申请 2021 年度国家自然科学基金优秀青年科学基金。持续三年与腾讯公司合作，利用人在回路的方法有效地融合了微信生态中用户的海量多源行为数据，提升朋友圈广告的点赞、评论等互动率 10%，提升广告主复投率 9%，取得了很好的效果。与新闻学院合作，探索多源数

据融合分析在社会公益领域的价值。例如：在武汉市新冠疫情爆发的早期，第一时间利用人在回路的数据融合技术高质地融合新浪微博、百度地图、社区、医院等数据，构建新冠肺炎求助者画像，刻画求助者的人口社会属性、健康状况，以及武汉市医疗资源的匹配程度，产生了较为一定的社会影响。

（2）荣誉获奖：

实验室联合北京大学、清华大学、上海交通大学、阿里云计算有限公司、京东商城（北京）数字科技有限公司、恒瑞通（福建）信息技术有限公司、上海数据交易中心有限公司等 8 家单位，承担国家重点研发计划“物联网与智慧城市关键技术及示范”重点专项项目《面向城市智能服务的数据治理体系与共享平台》，由中国人民大学教授杜小勇主持，总经费共计 3589 万元，其中国家专项资金 1489 万元，项目实施周期为 3 年。

（3）应用案例：

实验室依托技术创新联合体，从理论层面，结合前期智慧城市建设的实践，总结凝练出城市数据治理方法论体系并提供治理工具。从系统层面，围绕区块链系统为数据共享和服务提供可信任的基础，突破数据服务智能性不足以及区块链系统性能不高的瓶颈，研制数据服务智能化平台和可信数据共享平台。通过软硬件协同提升区块链系统性能的技术，在长三角国家数字经济试验区的大城市和国家级城市群开展跨行业跨区域的应用示范，验证理论与系统的有效性。

此外，实验室牵头 / 参与制订了国家标准“GB/T 34950-2017 非结构化数据管理系统参考模型”等 13 项。其中，《数据管理能力成熟度评估模型》被工信部明确选为待推广的重点国家标准。