Advances in Education 教育进展, 2026, 16(1), 133-138
Published Online January 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.161019

Hans汉斯

教育大数据架构分析与优化方案

常　文，张　雪，张海南

教育部教育技术与资源发展中心(中央电化教育馆)，北京

收稿日期：2025年11月30日；录用日期：2025年12月27日；发布日期：2026年1月4日

摘　要

本文围绕教育大数据架构及其优化方案展开研究。在分析教育大数据的概念、特点与发展现状的基础上，探讨了当前教育大数据建设面临的主要挑战。文章重点研究了教育大数据的架构体系，包括数据采集、存储、处理与分析等关键环节。通过对现有架构的深入剖析，提出了相应的优化策略，以推动教育大数据更加有效地支撑教育现代化进程，发挥其长远的积极作用。

关键词

教育大数据，教育数字化，教育大数据架构

Analysis and Optimization Plan for Education Big Data Framework

Wen Chang, Xue Zhang, Hainan Zhang

Center for Educational Technology and Resource Development, Ministry of Education, P. R. China (National Center for Educational Technology, NCET), Beijing

Received: November 30, 2025; accepted: December 27, 2025; published: January 4, 2026

Abstract

This paper focuses on the education big data architecture and its optimization scheme. Based on the analysis of the concept, characteristics and development status of education big data, this paper discusses the main challenges faced by the current construction of education big data. This paper focuses on the framework of education big data, including data collection, storage, processing and analysis. Through in-depth analysis of the existing framework, the corresponding optimization strategies are proposed to promote the education big data to more effectively support the process of education modernization and play its long-term positive role.

文章引用: 常文, 张雪, 张海南. 教育大数据架构分析与优化方案[J]. 教育进展, 2026, 16(1): 133-138.
DOI: 10.12677/ae.2026.161019

Keywords

Education Big Data, Education Digitalization, Education Big Data Framework

1. 引言

随着信息技术的快速发展和教育信息化的深入推进，教育大数据已成为推动教育变革和创新的重要力量。教育大数据的应用不仅能够提高教育教学质量和效率，还能为教育决策提供科学依据，促进教育公平和个性化发展。然而，当前教育大数据建设仍面临诸多挑战，如数据分散、标准不统一、存储水位增速快、系统架构不完善等问题。因此，深入研究教育大数据的现状和架构优化方案，对于推动教育大数据发展具有重要意义。本文将从教育大数据的概念和特征出发，分析其发展现状和面临的挑战，探讨架构优化的关键策略，通过设计多源异构数据采集方案、数据实时监测方案以及数据治理方案，推进教育大数据架构优化，实现数据统一汇聚、数据实时监控，并有效缓解数据存储压力，为充分挖掘数据要素价值、教育平台创新应用与长效发展提供有力支撑。

2. 教育大数据现状分析

2.1. 教育大数据的概念与特征

教育大数据的是指在教育活动中产生的大规模、多样化、高增长率和复杂性的数据集合。它具有四个典型特征：Volume (大量)、Variety (多样)、Velocity (高速)和 Value (价值)。教育大数据来源也十分广泛，包括在线教育平台、教学应用等。教育数据不仅包括传统的结构化数据，如学生成绩、考勤记录等，半结构化数据，如存在嵌套元素的结构化数据，还涵盖了大量的非结构化数据，如教学视频、在线讨论文本等。教育数据可以分为学习行为数据、教学管理数据等。学习行为数据记录了学生的学习过程，如学习时长、资源访问等；教学管理数据包括课程安排、教师评价等。通过对教育数据数据的汇集和深入分析，为教育研究和决策提供有力支撑。

2.2. 国内外教育大数据架构的介绍

随着在线教育行业的蓬勃发展，海量业务数据持续涌入数据存储系统，其规模已呈现指数级增长态势。传统关系型数据库在面对用户学习行为的多维度分析需求时，已显现出明显的性能瓶颈。这一现状促使数据仓库技术在在线教育领域获得前所未有的关注与应用。在数据仓库技术演进过程中，微软公司凭借其在全球软件行业的领先地位和强大的技术研发实力，在 SQL Server 2008 版本中实现了数据仓库技术的重大突破。该版本不仅提供了灵活的数据仓库扩展能力，还通过优化的接口设计和增强的数据转换服务，显著提升了数据整合效率与信息获取能力[1] [2]。与此同时，Oracle 公司作为数据仓库技术的先驱者，在 2012 年基于高效网络计算架构，创新性地开发了面向大规模数据处理的高性能数据仓库解决方案。该方案采用 X86 架构部署，在数据处理速度和分析效能方面实现了质的飞跃。特别值得一提的是，其集成的多维分析工具 Oracle Express 和 Oracle Discovery [3] [4]，通过先进的功能整合与需求响应机制，为数据存储与处理能力树立了新的行业标杆。而上述单独的数据仓库技术已经无法满足快速增长的大体

量数据的存储与计算需求，因此越来越多的在线教育大数据中心采用了基于虚拟化的大数据集群[5]进行建设，为数字教育平台的发展提供了重要支撑。

2.3. 国内教育大数据建设的现状及问题

2022 年，全国教育工作会议提出“实施教育数字化战略行动”。2025 年，中共中央、国务院印发了《教育强国建设规划纲要(2024~2035 年)》，为智慧教育平台建设指明了方向。各级教育机构和学校积极推进数据平台建设。国家智慧教育平台已汇集中小学资源 11 万余条[6]，职业教育在线精品课程 1.13 万余门[6]，高等教育优质在线课程 3.1 万门[6]，终身学习课程超 2000 门[6]，智教中国通行证累计注册用户突破 1.64 亿[6]，汇集了大量教育数据。而地方平台建设也取得突破，如重庆构建了“市－区县－学校”三级数字治理体系，完成学籍管理、教学资源等超万条核心数据的归集治理，数据合格率达 100% [7]。

为积极响应数字教育集成化发展的战略要求，构建高效的数据流通与利用体系，本方案提出教育大数据平台优化方案，致力于打破数据孤岛，实现教育数据资源的高效汇聚、整合与利用，为教育决策、教学创新和治理现代化提供坚实支撑。在推进这一目标的过程中，存在如下难题。一是数据标准化程度低，各级各类数字教育平台及工具的教育数据类型复杂多样，既包括账号、身份信息等结构化数据，也包含学习行为、教学互动等非结构化数据，在采集、存储和交换过程中缺乏统一规范，各类数据格式、编码规则和接口规范差异显著，导致数据共享互通困难。二是数据规模快速增长，随着教育信息化深入推进，各类平台产生的数据量呈指数级增长，部分地区资源有限，面对海量数据存储和实时计算要求，现有数据基础设施力不从心，存储扩容和计算能力面临较大压力。三是数据价值挖掘不足，当前多数教育大数据平台对数据的利用还停留在简单统计层面，对教学行为模式、学习效果关联分析等深层次的数据挖掘和应用，教育数据的潜在价值尚未得到充分释放。

为积极落实国家数字化战略行动中“集成化、智能化、国际化”的工作要求，充分释放教育数据潜在价值，基于大数据获取技术、大数据平台资源管理技术、大数据处理技术、企业级数据仓库[5]等，对数据本身的持续存储、清洗、分析、应用、管理和进化模式进行总结和凝练，形成成熟可复制的教育大数据整体解决方案，为教育信息化和教育大数据应用提供保障。

3. 教育大数据架构优化方案

为切实解决上述问题，本方案梳理了数据处理流程，覆盖数据采集、存储、计算、应用全过程，推进完善大数据集群技术架构和业务架构。通过构建多源异构数据采集、数据治理以及优化数据分析框架方案，切实解决教育数据流通中的关键瓶颈，为教育数字化高质量发展提供有力支撑。

3.1. 多源异构数据采集方案

本方案的设计主要为解决不同来源、不同结构的数据采集问题，促进多源异构数据共享、流通。

如图 1 所示，对于结构化数据(如 MySQL/Oracle 存储的教育业务数据)，采用 Apache Sqoop、DataX 等工具进行数据传输，将结构化数据抽取至 Hadoop 的分布式文件系统 HDFS、Hive 或 HBase 中；对于半结构化数据(如 JSON 日志、如用户行为等数据)，通过 Flume、Logstash 等日志管理系统工具、开发定制的数据采集 SDK，对数据收集、处理并写入数据接收方中。

3.2. 完善数据分析架构

针对当前教育大数据分析时效性问题，架构优化是关键。可以采用批处理和流处理相结合的混合架构，如图 2 所示。对于时效性强的数据分析需求，使用 Flink、Doris 等大数据处理框架，提高数据处理效

率。对于时效性不强的数据分析需求，可以使用基于 Spark on Hive、Tez on Hive、Spark 等计算引擎进行离线计算。同时，为保证数据安全，在数据的不同阶段可利用国密算法进行加密或散列等。在数据抽取落入存储中时，需通过存储加密层对文件存储加密、数据库敏感字段加密存储；在数据查询服务对外提供服务时，需经过传输加密层对加密或脱敏处理，以保证隐私信息不会泄露。

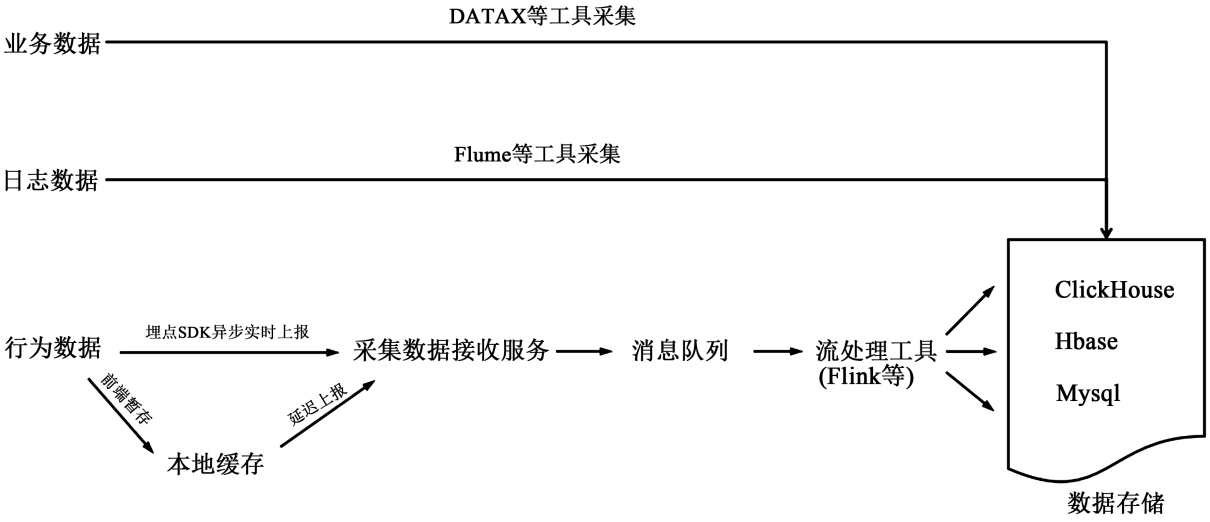

Figure 1. Diagram of multi-source heterogeneous data aggregation architecture
图 1. 多源异构数据汇聚架构图

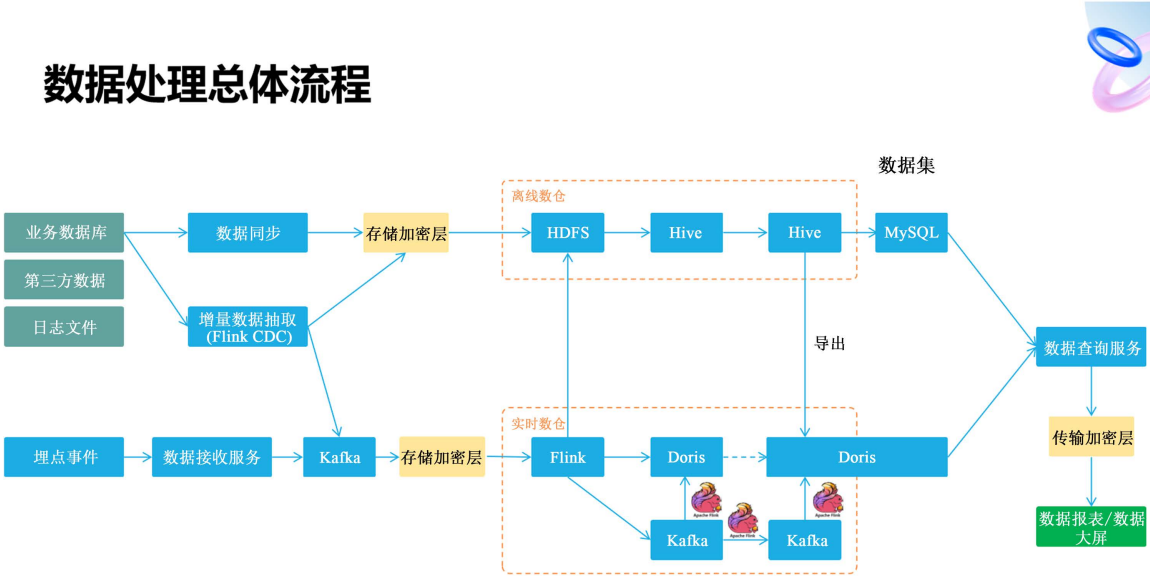

Figure 2. Diagram of the data analysis architecture
图 2. 数据分析架构图

3.3. 数据治理方案

数据治理方案为提高数据资产利用率，控制数据存储量级，优化计算和存储的性能并节省成本。一是深度梳理数据。对数据进行分级分类，研讨数据生命周期管理策略，并根据数据生命周期策略设计存算分离方案。二是实施存算分离方案。如图 3 所示，根据不同级别和类型的数据，采取不同的存算分离方式。

对于热数据(访问频率高的关键数据)使用高性能、低延迟的存储设备进行存储，如块存储。对于温数据(访问频率适中的重要数据)使用对象存储中性能匹配的产品进行存储，并利用大数据集群通过 Hadoop 分布式文件系统协议(即 HDFS 协议)读写。对于冷数据(访问频率低的非重要数据)归档迁移至低成本存储介质，如对象冷存储中的冷备存储中，以实现长期保存和优化存储效率的过程。三是建立数据质量管理体系。数据质量[8]建设是强调数据的准确性、完整性等基本特征。通过建立上报数据质量监测机制，引入数据治理机制，在数据入库前进行数据质量监测，对于不合格的数据不入仓管理，确保数据的准确性、一致性和时效性，进而使所有数据资源符合既定要求，去除低劣数据元素与数据单元，以此降低数据存储成本、增强数据分析效能。数据治理方案的实施可以实现教育大数据降本增效，为数智赋能打下良好的基础。

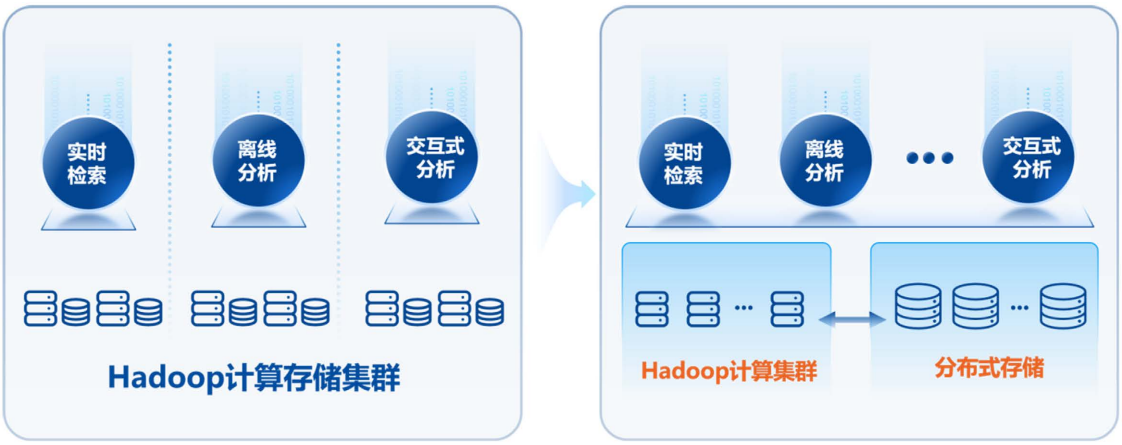

Figure 3. Diagram of the storage and computing separation architecture
图 3. 存算分离架构图

4. 结论

教育大数据建设是推动教育现代化的重要引擎。本文通过分析教育大数据的现状和挑战，探讨了架构优化的关键策略。为破解教育数据孤岛、数据存储效能低下等系统性难题，以“采、存、治、用”四维联动为框架，实现了教育数据从生产到价值转化的闭环管理，支撑教育大数据高质量发展。一是完善数据采集方案，将多源异构数据以多样方式进行采集；二是扩展数据分析架构，优化大数据框架，以多样大数据分析计算引擎提供不同时效性的数据分析服务，同时保障数据安全；三是推进数据治理方案，利用存算分离、数据质量监测等方案，将数据按不同业务情况进行存储或管理。通过大数据架构，推进各级各类在线教育平台和应用的数据交换，实现教育大数据数据分析赋能与降本增效。以准实时的教育分析数据，助力教育决策，赋能教学研管评各环节。研究表明，通过优化数据采集、存储、处理和分析等环节，可以有效提升教育大数据的应用价值。

然而，教育大数据建设仍面临诸多挑战，如数据安全、隐私保护、伦理规范等问题需要进一步研究和解决。未来研究应关注如何在充分利用教育大数据价值的同时，确保数据使用的合法性和伦理性。同时，还需要加强跨学科合作，推动教育大数据与教育学、心理学等学科的深度融合，为教育研究和实践提供更加丰富的视角和方法。

基金项目

本文系教育部教育技术与资源发展中心(中央电化教育馆)基本科研业务费专项资助“国家智慧教育平台体系建设技术路径与运行机制研究”(课题号 KZX202412)研究成果。

参考文献

[1] Barclay, T., Gray, J. and Slutz, D. (2000) Microsoft TerraServer: A Spatial Data Warehouse. *Proceedings of the* 2000 *ACM SIGMOD International Conference on Management of Data*, Dallas, 15-18 May 2000, 307-318. https://doi.org/10.1145/342009.335424

[2] 牛瑞瑞. 一种基于数据仓库的物流系统构建研究[J]. 信息与电脑(理论版), 2012(11): 37-38.

[3] Shahzad, M.A. (1999) Data Warehousing with Oracle. *Proceedings of SPIE—The International Society for Optical Engineering*, **3695**. https://doi.org/10.1117/12.339986

[4] Poess, M. and Othayoth, R.K. (2005) Large Scale Data Warehouses on Grid: Oracle Database 10 g and HP Proliant Servers.

[5] 周鹏. 基于大数据平台的 K12 在线教育数据仓库设计与实现[D]: [硕士学位论文]. 廊坊: 北华航天工业学院, 2020.

[6] 2025 年数字教育大会. 中国智慧教育白皮书[EB/OL]. https://wdec.smartedu.cn/doc/2025/《中国智慧教育白皮书》.pdf, 2025-05-16.

[7] 中国教育报. 重庆着力构建“智能驱动、产教融合、安全可控”的数字教育新生态——以数字化推动教育高质量发展[EB/OL]. https://baijiahao.baidu.com/s?id=1835158328253978561&wfr=spider&for=pc, 2025-06-17.

[8] 王正青, 但金凤. 大数据时代教育大数据治理架构与关键领域——以美国肯塔基州、华盛顿州与马里兰州为例[J]. 现代教育技术, 2019, 29(2): 5-11.