DOI：10.13955/j.yzyj.2022.02.08.04

中邮网院教育培训大数据分析与应用研究

刘树军，朱德军，柴立岩

（石家庄邮电职业技术学院，河北　石家庄　050021）

摘　要：以数据辅助领导决策、以数据驱动业务发展已成为企业数字化转型的必然选择。文章介绍了中邮网院大数据分析平台的系统架构、系统功能以及应用领域，探讨了中邮网院大数据分析应用的未来发展方向。

关键词：大数据；数据仓库；智能问答机器人；知识服务体系

中图分类号：F61　　**文献标识码：**A

2006 年 Hadoop 技术的出现标志着大数据技术时代的开始，10 多年来，大数据在宏观政策、技术创新、产业体系和应用场景等方面都得到了蓬勃发展。大数据的应用正在向全行业、全领域、全链条渗透，覆盖智慧政府、城市大脑、金融风控、健康医疗、疫情防控、精准营销等众多领域。

大数据时代，企业教育培训也走向了新的发展阶段，以“数据驱动业务发展”的数字化、精准化智能学习成为未来的发展趋势。本文探讨了作为中国邮政数字化培训平台，中国邮政网络学院（以下简称“中邮网院”）大数据分析平台的构建，旨在提供企业级一站式大数据采集、存储、计算、分析和应用的整体解决方案，并开展了相关的大数据和人工智能的应用研究和项目建设。

1　中邮网院大数据分析与处理平台架构

中邮网院大数据分析与处理平台集成数据采集、数据迁移、数据治理、数据存储、数据计算和数据应用等服务，涵盖了数据全生命周期，构建了大数据分析的技术路线，大数据分析平台架构如图 1 所示。

2　中邮网院大数据分析与处理平台功能

中邮网院大数据平台的核心价值体现在数据采得多、存得下、算得了、管得住和用得好，主要包含 3 个方面的功能：数据仓库、数据智能分析、数据可视化。

2.1　数据仓库

平台基于一站式大数据平台构建企业级数据仓库和数据集市，实现数据存储、检索、分析、计算，打通生产系统与大数据分析平台的数据连接，实现邮政业务、邮储银行等多板块，学员数据、培训数据、考试数据、学习数据和资源数据等多业务系统全量数据和增量数据的定时调度，可以进行海量数据的归集和存储。数据仓库开发的两个核心环节是数据迁移和任务调度。

基金项目：邮政应用技术协同创新中心资助项目（项目名称：中邮网院培训学习大数据分析应用研究；项目编号：YB2020004）

作者简介：刘树军（1983～），男，河北武安人，硕士，工程师，主要从事人力资源管理和大数据研究；朱德军（1984～），男，河北黄骅人，硕士，工程师，主要从事网络安全研究；柴立岩（1988～），男，河北馆陶人，硕士，工程师，主要从事软件开发研究。

收稿日期：2021—10—21　　**本刊网址：**yzyj.sjzpc.edu.cn

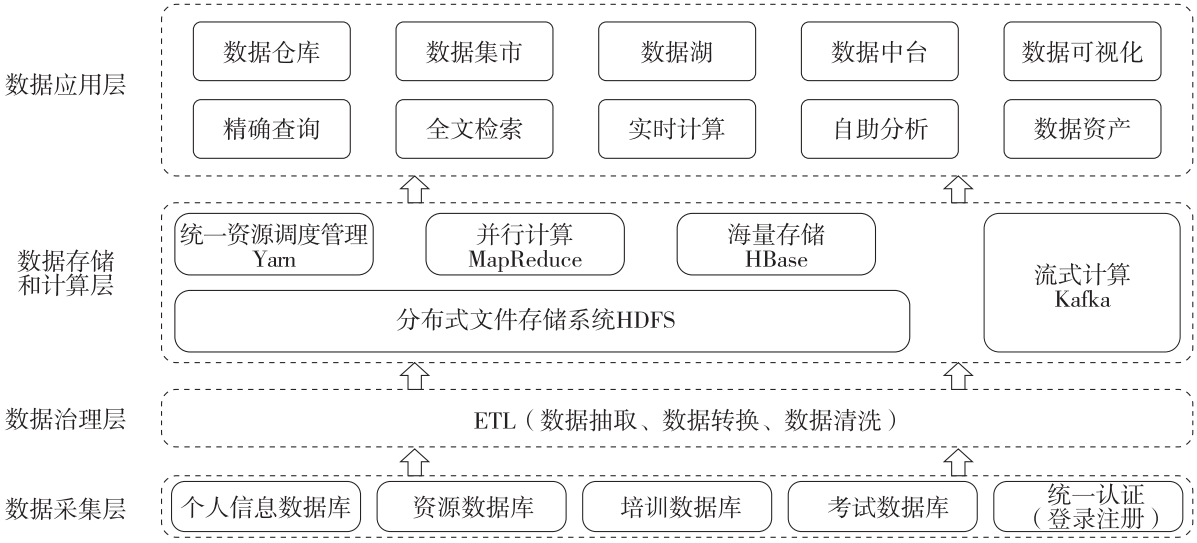

图 1　中邮网院大数据分析平台架构

2.1.1　*数据迁移*

生产系统产生的多源异构的业务数据，需要定时迁移到大数据分析平台，迁移策略分为批处理和流处理两种。

批处理类似于电梯中的直梯，当一部分人员进入直梯后将其运送到相应楼层。批处理包括全量更新和增量更新，全量更新是将生产系统数据一次性同步到大数据平台开展数据分析，适合数据量较小的表，一般为 100 万以内的数据量。增量更新是在一定周期内，根据数据变化将增量数据同步到大数据平台中。不同的源数据库增量同步策略不同，比如源端为 Oracle 数据库时，适合用 Oracle Golden Gate（简称 OGG）做增量更新，需要在源端和目标端同时进行 OGG 配置，原理是读取 Oracle 日志实现数据同步，增加、修改和删除等 DDL 操作的数据均可在目标端实现同步；源端为 MySQL 数据库时，适合用 Canal，原理是通过时间戳或标识列等自增长列实现增量日志更新，源端增加数据可以同步到目标端，而源端修改和删除数据时，不会在目标端体现。

流处理类似于电梯中的扶梯，一旦运行就不停止，一直处于运行状态，不断地运送人员。流处理需要分布式消息系统和流处理引擎结合使用，分布式消息系统一般采用 kafka，其优势在于实现与业务系统的解耦，同时具有消峰和抗压的作用，即数据量过大时可通过消息队列进行消费。开源的流处理引擎包括 Storm、Spark Streaming、Flink 等。

2.1.2　*任务调度*

针对不同的数据迁移方案，任务调度策略也分为两种。一种是通过调度时间配置，以及设置调度模式和依赖关系，实现周期性工作流自动调度。调度模式分为顺序调度和聚合调度：顺序调度是 A 工作流在调度周期内未执行完成时，在下个调度时间会按照顺序依次执行被延迟的调度；聚合调度是 A 工作流在调度周期内未执行完成时，在下个调度时间合并执行被延迟的调度。另一种是通过实时流程引擎，实现数据实时同步到大数据分析平台。

2.2　数据智能分析

支撑商业智能的大数据平台，融合事件驱动机制和复杂 SQL 编程模型的流处理引擎，具备全图形化的工具部署、运维和开发。提供数据导入、数据探索、数据预处理、特征工程、模型训练、性能验证、模型部署等全流程可视化建模能力，支持大规模分布式训练和自动化模型生成，实现不同业务场景的创建和管理，实现模型全生命周期管理，可以对海量数据进行大数据和人工智能分析与处理，有效挖掘数据中隐含的特点和规律。

2.3　数据可视化

中邮网院大数据平台支持 30 余种图表类型，可以将数据分析的结果以图形化的形式展示，提供强大敏捷的多维度分析功能，方便业务人员分析和使用。

3　中邮网院大数据分析的应用领域

3.1　智能问答机器人

针对邮政员工自助服务 App 中事务咨询模块问题重复率高、内容回复时间长、内容回复缺乏专业性和规范性的问题，中邮网院利用大数据技术，对事务咨询模块的问题进行聚类和语义分析，提取出共性问题，构建邮政员工人力资源知识库，并研发了智能问答机器人，通过中邮网院移动学习 App 向学员提供服务。

智能问答机器人的上线应用，有效提升了员工的使用便捷性，员工在日常工作中遇到的难题、与员工切身利益相关的政策、与员工生产行为相关的制度等，都可以在知识服务体系中找到答案，答案形式不仅包括文字，还有图片、视频、音频和链接等非结构化数据，上线三个月，员工问题量从每月 500 条提升到 2 000 条。员工输入问题的形式更加多元化，不仅可以通过文字输入，还可以通过语音方式输入，系统通过语义分析匹配出最优答案，即时进行回复，回复效率提升数十倍。在平台交互方面，智能问答机器人集成了评分、满意、不满意、常见问题推荐等功能，在不明确用户真实意图的情况下，还可以通过多轮对话逐步挖掘需求进行答复，提升了用户体验。

3.2　基于大数据的智能化推荐

随着远程培训资源在数量和规模上的不断扩大，网络上的资源呈爆炸式增长，员工在享受网络便利的同时，也受到信息过载和信息迷航的困扰，员工找不到最想要的课程资源，优质资源也无法触达员工，员工和学习资源之间没有形成有效的连接，导致员工能力脱节无法支撑企业高速发展。

中邮网院利用大数据分析技术，对 2 亿余条学习记录进行了分析，深入挖掘学员潜在的学习需求、学习规律和特征偏好，结合岗位能力要求，研发了智能化的推荐算法模型，解决了推荐模型的召回、选取、排序、冷启动等关键技术，构建了系统推荐、自主选择和新课推荐三位一体的推荐体系，最终在中邮网院移动学习 App 上线运行，基于学员真实的学习行为数据，实现了智能化资源推荐服务，有效提升了培训学习的个性化、科学化、针对性和体验性。

智能化推荐算法模型和推荐系统功能是根据企业培训实际，完全自主研发，具有可扩展性强、数据实时更新、用户操作简单等特点。系统提供了多维度推荐场景，并与业务深度融合，实现了员工学习需求、岗位能力要求和资源内容之间的智能化匹配，真正实现了以数据为导向的智能化学习，使得培训从“千人一面”转变为“千人千面”，员工所学即所需，提升了培训学习效能，赋能了员工职业发展。

智能化推荐系统面向邮政业务板块代理金融专业支局（所）经理和综合柜员 2 个岗位进行试点应用，覆盖人员 6.7 万人。智能化推荐系统的上线运行，一是有效解决了员工迫切反映的培训缺乏针对性、学习效率不高、学习体验不好等问题；二是有效连接了员工和学习资源，较大程度上避免了信息过载和信息迷航的困扰；三是将优质资源直达员工，促进了员工从被动培训向主动学习的转变。通过开展满意度调研，员工对推荐内容的整体满意度达 93.3%，系统的研发和上线为企业培训数字化转型进行了积极探索。

3.3　基于大数据的知识服务体系构建

目前中邮网院学员获取业务知识的主要手段是参加培训，学员的主动性不高，不能满足学员随时随地学习的需求。同时，培训内容还存在以下问题：知识资源主要是视频课件，呈现形式单调；知识载体主要以整个课件、整篇文档为单位向学员进行展示，而不是以知识点的形式存储，导致知识颗粒度较粗；知识以孤岛方式存在，知识之间没有建立联系，没有形成网状的知识体系等。

中邮网院利用大数据、人工智能和知识图谱等信息化技术，以问题为导向，通过 NLP 自然语言处理，对大量非结构化文本数据进行分句和分词处理，实现知识的抽取；通过图数据库技术，实现非结构化数据的存储。通过知识抽取、知识融合、知识推理和知识更新，构建基于实体—关系—属性的高级语义知识服务体系，实现知识推理、知识搜索、知识问答、知识推荐等功能，赋能邮政企业百万员工，满足员工对业务知识和管理知识实时获取的需要，为员工提供精准、立体的知识服务。

知识服务系统的构建，一是改变了传统的以培训为载体的知识灌输形式，构建了模式更新、粒

度更细、服务更智能的知识服务解决方案；二是改变了传统问答模式中问一答一的方式，答复的内容以网状形式进行关联；三是最大程度细化答复内容的颗粒度，以知识点形式进行呈现，而不是以整个课程的内容向用户进行展示，目前针对支局（所）经理和综合柜员 2 个岗位梳理了 15 000 余个知识点。知识服务体系的推广和利用，可以显著提高员工的工作效率，是知识传播的有力工具和企业文化传承的重要载体。

3.4　培训学习大数据分析

中邮网院在日常生产运营过程中，对培训学习开展情况进行数据统计是一项非常繁重的工作，有来自邮政集团人力资源部的要求，有来自邮储银行和中邮保险等各板块的需求，也有来自各省分公司对各省培训开展情况的诉求，而中邮网院培训系统的多样性和多态性，增加了企业信息化架构的复杂度。目前，邮政业务、邮储银行、中邮保险等各分院相互独立，各分院的业务培训、课程学习、考试竞赛、资格认证等后台数据又相互分隔，在企业内部形成诸多数据孤岛，每次数据统计分析工作都要从最底层的各业务数据库开始抽取数据，费时费力，无法高效支撑企业经营决策，也无法应对快速变化的业务发展，底层数据的互联互通成为困扰企业发展的痛点之一。

中邮网院利用大数据技术和平台，通过对培训学习产生的 12.6 亿海量数据进行采集、存储、计算、管理、挖掘和应用，整合邮政业务、邮储银行、中邮保险等各分院的多源数据，打破各系统的数据孤岛，进行可视化展示，转变为数据资产，快速形成数据服务能力。一是实现对学员培训学习行为的深入挖掘，发现学员最喜欢学习的时间集中在上午 9:00 ～ 11:00、下午 14:00 ～ 16:00 和晚上 19:00 ～ 21:00，全天 24 小时均有学员在学习，学习随时在发生。二是进行了全方位的业务运行情况分析，包括移动学习 App 运行情况、培训运行情况、课程学习情况、考试开办情况、直播运行情况等，年访问量超过 6 000 万人次，学习人次达 7 000 万，学习时长 1 500 万小时，年直播量 1 200 多场。三是了解用户群体特征，分析了各岗位群体、党员群体、三级领导等群体的培训开展情况，洞察不同群体的学习特点和规律，发现党员群体的学习积极性是非党员群体的 2 倍以上。四是构建了“人、培、学、考”多元化预警分析体系，使得管理人员在了解业务运行情况的前提下，根据预警分析结果及时发现业务开展过程中存在的问题和风险。通过多维度大数据的深度分析，为企业管理者提供经营决策服务，为企业精细化运营提供数据支撑。

4　中邮网院大数据分析的未来方向

目前，中邮网院大数据分析平台的应用主要集中在结构化数据的分析上，今后一段时间要将重点从结构化数据逐渐转向非结构化数据，结合人工智能算法，实现对图片、音频和视频等信息的分析，例如基于大量的音视频课件，利用深度学习对音视频进行场景分类、人物识别、语音识别、文字识别等多维度分析，形成层次化的分类标签，支撑准确高效的视频搜索，使用户根据关键词即可搜索出相关的音频和视频内容，内容可以具体定位到知识点，提升搜索体验。

中邮网院大数据分析平台可通过有效洞察数据价值，从多个维度挖掘新的业务增长点，分析学员学习行为，挖掘学员学习特点和规律，提供智能问答机器人、智能推荐、知识服务体系服务，以有效应对快速变化的业务发展，为管理者提供企业经营决策支持，赋能企业数字化转型。

参　考　文　献

［1］新华社．中共中央 国务院关于构建更加完善的要素市场化配置体制机制的意见［EB/OL］． http://www.gov.cn/zhengce/2020-04/09/content_5500622.htm，2020-04-09

［2］中国信息通信研究院．大数据白皮书（2020 年）［Z］．北京：中国信息通信研究院，2020

［3］大数据产业生态联盟，赛迪智库．2020 中国大数据产业发展白皮书［Z］．北京：大数据产业生态联盟，2020

［4］中国邮政．集团大数据平台整体方案项目概述［EB/OL］．https://max.book118.com/html/2020/1212/6233034153003033.shtm，2020-12-12

［5］钱智君．大数据技术在邮政企业中的应用［J］．电子技术与软件工程，2018（2）