文章编号：1005-8451（2016）09-0001-06

铁路大数据平台总体方案及关键技术研究

史天运[1]，刘　军[2]，李　平[2]，马小宁[2]
（1.中国铁道科学研究院　电子计算技术研究所，北京　100081；
2.中国铁道科学研究院　铁路大数据研究与应用创新中心，北京　100081）

摘　要：大数据是当今炙手可热的技术词汇，在全球掀起一场思维变革，将成为新一轮科技和产业竞争的前沿。大数据技术对于提升中国铁路总公司核心竞争力及推动铁路转型升级都具有不可估量的作用。本文阐述了铁路大数据的基本概念和特点，分析了铁路大数据应用的现状及需求，设计了铁路大数据平台的总体架构，剖析了铁路大数据应用的关键技术。对促进大数据技术在铁路行业的应用研究具有一定的指导意义。

关键词：铁路；大数据平台；Hadoop
中图分类号：U29-39　　**文献标识码：**A

Overall scheme and key technologies of big data platform for China Railway

SHI Tianyun[1], LIU Jun[2], LI Ping[2], MA Xiaoning[2]
(1. Institute of Computing Technologies, China Academy of Railway Sciences, Beijing　100081, China;
2. Research and Application Innovation Center for Big Data Technology in Railway, China Academy of Railway Sciences, Beijing　100081, China)

Abstract: Big data, which is today's hottest technical vocabulary, is setting off a global thinking change, and become the forefront of a new round of technological and industrial competition. Big data technology has an immeasurable role to enhance the core competitiveness of China Railway and promote the transformation and upgrading of the railway. This paper expounded the basic concept and characteristics of railway big data, analyzed the current situation and demand of railway big data application, designed the overall architecture of railway big data platform, analyzed the key technologies, provided some guidances to promote the implementation and application of big data technology in China Railway.

Key words: railway; big data platform; Hadoop

由于智能传感器的广泛应用及信息技术的迅猛发展，人类产生并存储的数据量呈爆炸式增长，数据在人类生产、生活中扮演着越来越重要的角色，大数据在此背景下应运而生。近年来，美、欧、日、韩等发达国家纷纷制定大数据国家战略，加快大数据布局。我国从2015年开始也颁布了《促进大数据发展行动纲要》[1]、《互联网＋行动计划》等一系列文件，并在互联网、交通、电信、金融、电力、征信等领域积极开展大数据应用示范。目前，全球都正处在一个思维变革、数据创新的浪潮之中。

随着高速铁路的快速发展及铁路信息化建设的逐步深入，中国铁路已积累了海量的结构化、半结构化、非结构化的数据，包括12306网站和95306网站的客、货运数据，设备台账数据，基础设施检测数据，自然灾害监测数据，视频监控数据和工程建设图纸等。据初步统计，铁路总公司以及各铁路局存储的数据总量已达到10 PB的数量级，且各类数据增量极快，大量视频图片仅保存极短时间。可以说，中国铁路已步入大数据时代。

大数据技术[2~5]在铁路的应用，不仅有利于促进数据资源共享，盘活铁路数据资产，探索新的利益增长点，更有助于保障铁路行车安全，提升铁路服务水平，增加铁路企业的经济效益[6~9]。现阶段，急需总结铁路数据资源的现状及存在问题，明确铁路各业务领域对大数据的需求，强化顶层设计，突破核心技术，在典型领域开展大数据应用示范，以应用促进大数据在铁路的应用研究。

收稿日期：2016-06-15
基金项目：中国铁道科学研究院重大课题（2015YJ080）；中国铁路总公司科技研究开发计划重点课题（2015X003-B，2015X003-C，2015-X003-F）。
作者简介：史天运，研究员；刘　军，助理研究员。

1 铁路大数据的概念及特征

1.1 基本概念

对于大数据技术的概念，目前还没有形成一个公认的提法。许多公司都从自己的角度进行解读，以下列举几个典型的提法。

维基百科：大数据是指一个超大的、难以用现有的数据库管理技术和工具处理的数据集。

麦肯锡：大数据是指无法在可承受的时间范围内，使用常规软件工具进行采集、捕捉、管理、处理的数据集合。

Gartner：大数据是指需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

Informatica：大数据是指涉及交易和交互数据集在内的所有数据集，其规模或复杂程度超出了常用技术可按照合理的成本和时间采集、存储、管理及处理这些数据集的能力。

铁路大数据是大数据技术在铁路行业的缩影，是指由铁路客运、物流、基础设施、移动设备、工程建设、资产经营、企业管理等各业务领域的结构化、非结构化数据所汇集而成的数据集合。数据量大，数据类型多，需要通过新型大数据技术的应用才能快速开展数据的采集、抽取、存储、检索、分析、挖掘和展示，并从大量的数据中挖掘出隐藏的业务规律、发展趋势，最终达到提高运输组织效率、保障铁路行车安全、优化客货服务质量、提升企业经营效益的目的。

1.2 典型特征

铁路大数据的典型特征是数据量大、数据类型多、数据增长快、业务价值大。

（1）数据量大

近年来，随着铁路信息化建设的逐步深入，信息系统已覆盖客货营销、运输组织、经营管理等各个领域，各系统都积累了海量的数据。特别是，随着12306网站及95306网站的上线，售票信息及铁路物流信息大幅增长。基础设施及设备检测方面，铁路的工务、电务、供电、车辆和机务等部门积累了铁路线路、通信信号、机车车辆等各种设施设备的海量实时状态数据。

（2）数据类型多

铁路行业数据主要包括结构化数据、非结构化数据和流数据。结构化数据主要包括：业务系统的基础数据、业务数据、统计汇总数据等。非结构化数据主要包括：沿线和车站监控视频、铁路工程建设图纸和设计文档、语音服务数据和办公文件等。流数据主要包括：设施设备实时检测数据、铁路机车设备实时状态数据、列车实时控制数据等。

（3）数据增长快

目前，高速铁路沿线布置的摄像头、检测设备、控制设备等，每天产生大量的非结构化数据，且增量巨大。12306网站和95306网站每天都会产生大量的订单数据和网上购票行为数据。

（4）业务价值大

铁路行业数据资源具有巨大的应用潜力及价值。如，铁路售票数据对于精准营销、优化开行方案、联合出行规划具有重要的意义。铁路物流数据对于优化物流流程、提升物流的精细化水平也具有重要的作用。

2 铁路大数据应用现状及需求

2.1 铁路大数据应用现状

（1）虽然铁路信息系统建设近年来逐步完善，但各系统各自为政，独立建设，数据共享备份不够，集成较弱，特别是基础数据多头维护，统一管理需加强。

（2）数据管控力度薄弱，数据标准化程度不高，存在数据不一致、不准确问题，数据质量有待提高。

（3）铁路总公司、铁路局和站段之间网络带宽相对不足，数据采集的及时性无法保证，各级系统间的数据交换难以实现。

（4）技术手段薄弱，仍采用传统的数据库技术、数据处理技术开展大数据的应用分析，缺乏专用技术及工具支撑，数据处理的时效性、可用性不强。

（5）对于数据的利用还停留在初级阶段，深层次的数据分析、数据挖掘较少；同时，对于数据的利用仍以专业为界限，缺乏跨部门、跨业务系统之间的数据综合分析。

（6）铁路数据共享模式不成熟。为实现数据综合分析，需采集不同业务系统的数据，但不同部门在合作模式不清晰情况下，不愿意提供铁路业务数据，需先解决不同业务部门之间合作的“共赢”模式。

2.2 铁路大数据应用需求

（1）总体数据规划，统一数据标准

突破部门及单个信息系统的界限，立足整个铁路行业，从整体、宏观的角度理清数据流程，明确数据资源的分类、分级及数据模型，推进数据资源的标准化，包括元数据、数据元及数据库等标准，促进系统间的互联互通。

（2）强化数据治理，提升数据质量

从组织、流程、技术等不同维度出发，构建完善的数据治理能力。建立数据管理维护组织，明确数据生产者、维护者、使用者等的责权，建立标准化的数据管理维护流程，建立数据评价考核指标体系，健全数据治理工具，最终达到提升数据质量的目的。

（3）加强数据开放，促进综合应用

大数据一个突出的亮点就在于实现跨领域的综合分析。应转变理念，打破部门鸿沟，树立开放共享的思想，促进系统的互联互通，从而实现数据的共建共用、融合创新。

（4）加强数据清洗，确保源头质量

在数据资源进入大数据平台之前，需要大力开展数据清洗工作，及时发现并解决数据质量问题。需识别并删除重复数据，补充缺失值，光滑噪声数据，确保数据的唯一性、准确性、完整性。

（5）强化基础设施，提升处理能力

在硬件设施方面，首先需要升级既有的服务器、存储设备，具备大数据分析基础的物理条件。同时，基于先进的分布式存储、分布式计算、流计算、内存计算等技术，搭建大数据分析处理基础架构，提供 PB 级数据的离线计算能力，以及 TB 级数据的实时计算分析能力，支撑各业务领域开展大数据分析工作。

（6）完善体制机制，保障数据安全

数据开放共享意味着数据面临更大的安全威胁。铁路大数据平台中存储和处理不同安全级别的数据，需从机构、管理、技术多方面统筹考虑，构建完善的数据安全保障体系，防止数据被窃取，数据被非法修改、非法复制等。

（7）明确应用场景，深挖数据潜能

大数据本身不具备价值，必须和具体的应用场景相结合才能发挥作用。铁路大数据应用，重中之重还是基于行业特色，厘清业务痛点，结合业务发展趋势，找出大数据与业务的结合点，明确大数据分析的应用场景。未来可在铁路客运、货运、基础设施检测、动车组、运输安全等领域开展数据分析和挖掘，为领导提供重大决策的支撑信息，挖掘新的业务增长点；为各业务部门提供跨部门的有价值信息，提升在铁路领域中的核心竞争力。

针对客运领域，需开展发送客流和始发站—终点站（OD）客流的长短期预测，并对客户按价值进行分群，建立旅客积分、奖励制度，提高旅客满意度和忠诚度，吸引和稳定客户资源。同时，提供送票、餐饮、酒店、旅游、租车的旅客个性化推荐服务，构建高品质、多层次、全方位、立体化的铁路客运服务。对网络黄牛、抢票软件用户、网络爬虫等异常用户的行为数据进行智能识别。

针对货运领域，需开展铁路货运量预测与预警，货运客户分级评价与流失预警和铁路货运物流优化。

针对运力资源，需实时监控不同基础设施设备、机车车辆设备、环境监测设备，对设备状态进行科学地评估、对运行故障准确地诊断、发现设备状态全生命周期的演变规律，对服役状态进行预测分析，延长设备的使用寿命，高效指导养护维修。

针对动车组管理，需实时监测动车组各种状态，保障动车组运行安全，降低动车组运用维修成本，优化动车组及维修资源配置，提高动车组维修效率。

针对铁路运输安全，需构建全面、全员、全过程的安全风险控制体系，重点开展关键部件失效规律及模式之间的关联性分析，实时监控机车、动车组、客车、货车等移动设备，保障铁路运输安全。

3 铁路大数据平台总体方案

3.1 建设目标

在铁路信息化总体规划的指导下，以强化数据治理、提升数据质量为基础，以共享交换、挖掘分

析及精细化管控为目标，以基础数据的规范统一、集中管理为支撑，统一数据标准，实现数据资源的充分共享及综合应用；搭建基于数据仓库、大规模并行处理（MPP，massively parallel processing）和Hadoop的综合性大数据分析基础架构，提供离线的PB级和在线的TB级数据处理能力，支撑各业务领域开展深入的数据挖掘分析及业务模式创新；及时全面地掌握数据资产的内容、存量、增量及使用情况，实现对于数据资产的全生命周期精细化管控；为提高运输组织效率、保障铁路行车安全、优化客货服务质量、提升企业经营效益提供支撑。

3.2 建设内容

3.2.1 铁路基础数据管理平台

统一、准确的基础数据是开展大数据分析的基本前提。为改变铁路各业务系统基础数据的“自采集、自管理、自维护”局面，消除基础数据的不一致、不完整、不及时等问题，铁路基础数据管理平台实现基础数据的统一管理、共建共用，对基础数据进行数据清洗、数据标准化，提供权威、及时、全面的基础数据。

3.2.2 铁路数据服务平台

在梳理铁路业务系统数据的基础上，建立铁路行业数据目录全视图，确定铁路数据的保密级别，建立铁路数据共享的渠道，完善铁路数据共享的审批机制，形成客运、货运、基础设施检测、联调联试、动车组、行车安全等的数据主题，探索铁路数据共享在管理和商业方面的“共赢”模式。

3.2.3 铁路大数据分析平台

搭建基于数据仓库、Hadoop和Spark等技术的铁路大数据分析平台，实现PB级离线和TB级在线数据处理能力，可处理传统关系型数据库数据、传感器的流数据、视频和语音的非结构化数据，支撑各业务领域的数据分析需求。

铁路大数据典型应用场景包括客运、货运、基础设施、运输安全等各领域。

（1）客运领域

运用大数据技术实现发送客流和OD客流的长短期预测，优化列车次数、开行数量和计划；开展铁路旅客网上购票行为分析，建立旅客全方位的用户画像，分析旅客满意度和忠诚度，并实现送票、保险、餐饮、酒店、旅游、租车的个性化智能推荐服务；对网络黄牛、抢票软件等异常行为实现智能预警；实时收集航空的票价信息，实现铁路商务舱票价智能调整。

（2）货运领域

运用大数据技术实现物流总量的长短期预测，建立物流货主的用户画像并进行客户细分，分析货主对铁路运输的粘合度，进行货主流失预警；实时收集航空、公路、水运等交通方式和其他物流公司的运价，辅助开展铁路货运运价的调整，以及提升铁路物流的服务能力和水平。

（3）运力资源领域

通过实时采集工务专业的TQI、高低、轨向、轨距、水平、三角坑等数据，电务专业的机车掉码、电化干扰、邻线干扰、邻段干扰、场强覆盖、服务质量等数据，供电专业的超限数据和视频数据，车辆专业的测温精度、探测角度、轴距、无线数传等数据，以及机车和环境监测设备的状态数据，实现基础设施、机车等设备状态科学评估，分析设备故障演变规律，优化设备维修计划。

（4）运输安全领域

运用深度学习实现基础设施和移动装备及关键部件的故障诊断和识别，实现基于多数据源融合的基础设施和移动装备故障智能识别，分析基础设施和移动装备运行总趟数、风、雨、雪等环境因素与设备故障率的关联关系。

3.2.4 大数据分析模型库

建立并维护预测、分类、聚类、关联分析、支持向量机、神经网络、优化方法、智能推荐等基本算法库，以及工务、电务、供电、车辆、机务等专业基本模型库。

3.2.5 铁路大数据可视化平台

提供完整的、统一的、多维度、多层次的数据可视化展现能力，包括联动的、动态的、二维的、三维的柱状图、散点图、饼图、雷达图、地图、仪表盘、图谱、GIS等。

3.3 总体架构

铁路大数据平台总体架构主要由数据采集层、数据传输层、数据存储层、数据服务层、数据分析层、

数据应用层、数据展示层及数据标准体系、数据保障体系组成，如图 1 所示。

图1　铁路大数据平台总体架构

（1）数据采集层

铁路大数据主要包括路内数据和路外数据。路内数据主要涉及客运、物流、工程建设、基础设施检测、动车组管理、联调联试、营销、客户、运输组织、运力资源、安全监控、财务、人力、物资等，路外数据主要涉及航空、公交、地铁、出租车、旅游、酒店、气象、互联网等。

（2）数据传输层

数据传输层包括铁路内网、铁路专网、4G、LTE、GSM 和 WIFI，实现铁路总公司—铁路局—站段—终端设备之间的快速数据传输。

（3）数据存储层

现有铁路业务系统中的数据大多以关系型数据库进行存储，包括 Oracle、MySQL 等；对于非结构化数据通过 Kafka、Sqoop、Flume 软件，将数据转换成为文件方式进行存储。

（4）数据服务层

数据服务层主要包括数据治理、数据共享和数据主题域。数据治理包括数据清洗、主数据管理、数据标准化和数据质量管理；数据共享服务包括铁路数据目录服务和脱敏数据样列；数据主题域主要包括客运、货运、基础设施、动车组、联调联试等主题数据。

（5）数据分析层

数据分析层分为结构化数据分析和非结构化数据分析。结构化分析基于数据仓库和 MPP，非结构化分析基于 Hadoop、Spark、Storm、Hive、Mahout、R 等。铁路大数据分析算法主要包括回归分析、分类、聚类、关联分析、统计分析、主成分分析、支持向量机、神经网络、深度学习、优化算法和智能推荐等基础算法。

（6）数据应用层

通过铁路大数据分析平台的挖掘结果，实现对客运领域、货运（物流）领域、基础设施领域、动车组领域、联调联试领域的主要业务提供强有力支撑，并开展新业务的挖掘。

（7）数据展示层

通过折线图、柱状图、散点图、饼图、雷达图、地图、仪表盘、漏斗图、树图、标签云、图谱、GIS、3D 进行数据展示，以及二维和三维可视化区域的扩大、缩小和移动等，实现铁路数据动态的、实时的、联动的展示，可重点开展基于 GIS 平台的铁路客运迁徙图，以及铁路客户之间的关联图等。

（8）标准体系

铁路大数据标准体系主要包括铁路大数据采集、存储、管理、共享、使用、安全等标准规范，明确数据的范围和格式、数据管理的权限和内容、以及

数据接口对接格式和访问方式等。

（9）保障体系

铁路大数据平台保障体系主要包括数据及网络信息安全保障、运行维护保障、人才技术保障、评价考核保障。

4 铁路大数据关键技术

4.1 基于博弈论的铁路数据共享技术

铁路数据共享问题是一个博弈论利益分配的问题，即共享数据产生价值和数据责任最优分配问题。数据共享多者收益越大，所承担责任多者收益越大，以及数据价值越大提供方收益越大。建立基于博弈论的数据价值最优分配方式，使数据提供方、技术支持方、数据经营方在商业上达到“多赢”。

4.2 铁路数据仓库、Hadoop和Spark混合技术

针对不同类型数据处理需求，结合业界主流做法，铁路大数据平台采取混搭架构，由数据仓库、MPP 和 Hadoop 构成。数据仓库主要负责高性能数据加工及综合分析等企业关键应用，存储最核心的跨域业务数据；MPP 负责存储长周期历史数据，进行深度分析及自助分析应用；Hadoop 主要存储海量非结构化数据，进行相关的批量处理及探索挖掘分析。

4.3 铁路大数据分析算法和专业模型

铁路大数据分析算法包括回归、分类、聚类、关联分析、统计分析、主成分分析、支持向量机、神经网络、深度学习、优化算法等核心机器学习算法。专业模型包括铁路客运、货运、基础设施、营销、经营、动车组、联调联试等各专业用于分析或预测的专业模型。

4.4 大规模参数的机器学习优化技术

铁路大数据分析平台提供神经网络、深度学习、支持向量机、分类、聚类、智能推荐等多种机器学习模型，涉及大量参数学习和参数优化。目前，机器学习参数优化方法是基于梯度下降的优化方法，具有陷入局部极值的不足，需研究基于并行的、实时的、不易陷入局部极值的优化方法，快速调整模型中关键参数，使模型在较短时间内具有较好的效果。

4.5 基于深度学习的铁路视频和图像分析技术

随着高速铁路的迅猛发展，在高速铁路沿线、车站、高速列车内、综合检测车等安装了大量视频摄像头，各种视频中包含了大量涉及铁路安全、资产经营等方面的信息。需通过深度学习算法逐层提取图像的基本特征，为设备故障自动识别、基础设施自动检测提供强有力的技术手段，保证铁路列车运行安全，提高铁路基础设施检测效率。

5 结束语

大数据技术是当今最为活跃的新技术，是促进业务创新增值、提升企业核心价值的重要驱动力。大数据技术与铁路的结合具有深刻的现实意义，在客户画像、市场营销、产品设计、行车安全、服务质量、设备管理等各个方面都将发挥显著的作用。本文对促进大数据技术在铁路行业的应用研究具有一定的指导作用。

参考文献：

[1] 中华人民共和国国务院．国发（2015）50号 促进大数据发展行动纲要 [Z]. 北京：中华人民共和国国务院，2015，8.

[2] Apache Hadoop[EB/OL]. https://en.wikipedia.org/wiki/Apache_Hadoop, 2015.

[3] Big Data[EB/OL]. https://en.wikipedia.org/wiki/Big_data, 2016.

[4] Apache Spark[EB/OL]. https://en.wikipedia.org/wiki/Apache_Spark, 2016.

[5] MapReduce[EB/OL]. https://en.wikipedia.org/wiki/MapReduce, 2016.

[6] 中国铁道科学研究院．大数据技术在铁路行业中的应用研究 [R]. 北京：中国铁道科学研究院，2015，6.

[7] 中国铁道科学研究院．云计算技术在铁路行业中的应用研究 [R]. 北京：中国铁道科学研究院，2015，6.

[8] 中国铁道科学研究院．物联网技术在铁路行业中的应用研究 [R]. 北京：中国铁道科学研究院，2015，6.

[9] 中国铁道科学研究院．北斗卫星导航技术在铁路行业中的应用研究 [R]. 北京：中国铁道科学研究院，2015，6.

[10] Jun Liu, Tianyun Shi, Ping Li. Optimal Cloud Storage Problem in the Distributed Cloud Data Centers by the Discrete PSO Algorithm[C]. 2015 IEEE Conference on Evolutionary Computation, pp.156-163, 2015.

[11] Jun Liu, Ping Li,Tianyun Shi and Xiaoning Ma. Optimal Site Selection of China Railway Data Centers by the PSO algorithm[C].12th World Congress on Intelligent Control and Automation, pp.251-257, 2016.

责任编辑 王 浩