文章编号：1005-8451（2016）09-0022-03

基于大数据平台的铁路客运数据分析技术方向研究

王洪业，王炜炜，贾欣茹，单杏花
（中国铁道科学研究院　电子计算技术研究所，北京　100081）

摘　要：数据本身蕴含着价值，对于已经产生的数据，其价值量会随着时间的推移逐渐减少，甚至会因为存储数据而要付出不菲的成本。如何利用相关分析技术和工具把数据中蕴含的价值挖掘出来，是每一个企业都要解决的一个难题。本文在介绍铁路客运数据仓库及分析技术的基础上，通过对基于大数据平台的分析技术方向进行研究，提出了适合铁路客运的分析技术路线。

关键词：大数据；数据分析；数据仓库；企业分析数据集
中图分类号：U293：TP39　**文献标识码：**A

Railway passenger transport data analysis techniques based on big data platform

WANG Hongye, WANG Weiwei, JIA Xinru, SHAN Xinghua
(Institute of Computing Technologies, China Academy of Railway Sciences, Beijing　100081, China)

Abstract: The data itself contains a value, for the generated data, its value will be gradually reduced over time, and even to pay the expensive cost because of the stored data. How to use relevant analytical techniques and tools to excavate the value contained in the data, is a puzzle to be solved for every enterprise. Based on the introduction of data warehousing and analysis techniques of railway passenger transport, and through the research based on the technical direction of big data platform, this article put forward a technical line for railway passenger transport data analysis.

Key words: big data; data analysis; data warehousing; business analytical data set

当今社会是数据的社会，数据无处不在，并且在急剧地爆炸式增长，大数据已经走进了我们的生活、企业的生产和社会的发展。面对大数据的激流、多元化数据的大量涌现，大数据已经为个人生活、企业经营，甚至国家和社会都带来了机遇和影响[1]。

大数据的数据量很大，但大数据只是数据，自身蕴含着价值但不会自动产生价值；要想挖掘出大数据所蕴含的价值，必须通过数据分析和挖掘，得到切实可行的行动建议、方案等，用这些行动建议和方案指导生产、创造出价值。一个企业拥有多少数据并不重要，重要的是能够通过使用创新的、强大的分析方法，从数据中汲取价值。

随着我国铁路客运数据的累积以及客运规模的迅速扩大，客运数据种类也越来越多，历史数据规模越来越大，仅仅依靠关系型数据库进行数据的管理和操作，已经不能满足数据分析的要求，制约了从客运大数据中挖掘出价值的步伐。为了能够跟得上大数据发展的步伐，铁路客票系统技术团队通过大数据技术构建了铁路客运数据分析平台，采用 Hadoop 体系结构，实现了对海量数据的操作、管理和应用展现。

1　铁路客运数据仓库及分析技术

随着客票系统的建设与发展，铁路客运营销辅助决策系统也经历了十几年的发展，建成了铁路总公司级、铁路局级的营销系统。经过多年的积累，铁路总公司级营销数据仓库已成为国内最大、最全面的铁路客运信息存储、应用中心，采用了 Sybase IQ 数据库产品，管理着客票交易、席位、订票以及相关业务数据的历史记录，并提供多种分析主题、数据集市的应用。

收稿日期：2016-06-15
基金项目：国家自然科学基金（U1334207）；中国铁路总公司科技研究开发计划课题（2016X005-B）；中国铁路总公司科研计划课题（2016-X004-G）；中国铁道科学研究院科研专项课题（研发中心）（J2016X009）。
作者简介：王洪业，副研究员；王炜炜，副研究员。

但是随着客运历史数据的累积，以及全国铁路客运规模的快速扩展，铁路总公司级数据仓库的规模已经达到几个 TB，在数据传输、加载、查询的效率方面逐渐下降，尤其是对数据量较大的表进行关联查询，需要几十分钟甚至数小时，影响了用户统计分析的要求。

铁路客运数据分析平台的建设思路是作为现有营销系统平台的补充，基于目前的铁路客运营销数据仓库体系，有机结合 Hadoop 大数据处理体系，形成能力更加强大、规模可以扩展、应用更加广泛的营销数据基础平台。

铁路客票数据分析平台是铁路客运数据分析平台的一部分，它由 Sybase 数据仓库平台与 Hadoop 大数据平台结合而成，其体系结构如图 1 所示。

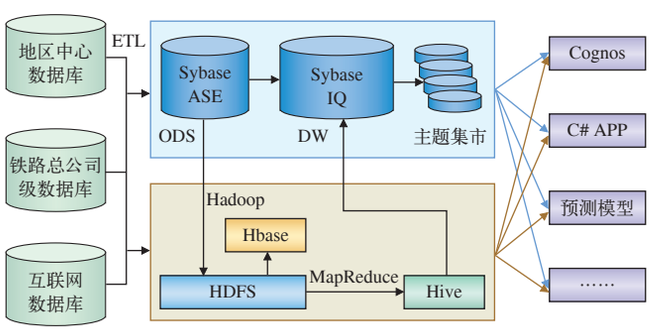

图1　铁路客票数据分析平台

铁路客运数据分析主要是通过提供工具建立数据挖掘、分析模型，进行旅客行为、运营效益、客流预测等业务专题的分析挖掘。

（1）客流预测建模功能

提供工具建立客流预测模型，进行发送总量、始发站—终点站（OD）客流、趟车等专题预测。

（2）客运业务专题分析功能

铁路客运大数据分析在以下几个方面需要进一步加强：

a. 铁路客运数据分析更多的是对数据进行统计并生成报表，是对趋势进行分析。分析重点还停留在大数据分析的初级和中级阶段，对智能引擎、预测模型、行动建议等涉及不多。

b. 铁路客运数据分析的数据源只占铁路客运所产生数据的一小部分。目前分析的重点数据主要是客票交易时所产生的结构化的数据，规模更加庞大的非结构化的数据并没有真正地应用于数据分析过程中。

c. 数据收集的范围有待进一步提升。目前应用最为广泛且为人们所熟知的大数据源应该是从互联网上收集来的、用来记录用户操作日志的数据。用户浏览互联网所产生的日志信息，蕴含着很高的价值，是等待分析和挖掘的资源宝库。通过 12306 网站收集用户操作的所有日志信息，为分析挖掘提供坚实的基础。

2　基于大数据平台的分析技术方向

基于大数据平台的分析技术方向必须符合大数据思维，要运用大数据思维去分析、发掘潜在的价值。大数据思维主要包括：需要全部数据样本而不是抽样、关注效率而不是精确度、关注相关性而不是因果关系。发展是时代永恒的主题，数据在变化、分析技术在发展，在数据分析的世界里，使用新的数据源，新的数据分析技术，突破当前分析瓶颈，挖掘出更大的价值是数据分析发展的方向。基于大数据平台的分析技术方向应该包括分析方法的可扩展性、创新性和数据源的不断突破性。随着业务数据的不断增加，大数据分析的技术必须易于扩展并且要不断的扩展，大数据分析的方法要不断更新，用于分析的数据种类要不断丰富，这样才能适应发展的要求，才能在大数据分析领域与时俱进，才能分析挖掘出更大的价值。

在大数据分析领域，有几项重要的技术，使用好这几项技术，将会使大数据分析更具可扩展性，并带来分析方法的革命，分析、挖掘出更大的价值。这几项技术包括：云计算、海量并行处理架构（MPP，Massively Parallel Processing）和 MapReduce。

云计算具备以下 3 个重要特征 [2]：

（1）企业无需进行基础设施建设，没有固定资本的支出，有的只是运营成本；

（2）系统能力可以在很短的时间内显著地扩大或缩小；

（3）云计算的底层硬件可以在地理意义上的任何地方。云分为两种：公有云和私有云。

MPP 数据库会把数据切成不同的独立数据块，由独立存储和 CPU 资源进行管理，这样的优势是可

以对海量的数据进行切割并实现并行处理。此处所提的MPP，并不是上述意义上的MPP，而是使用类似MPP处理数据的方式方法进行数据分析和挖掘。

MapReduce是一种并行的编程架构，包括映射过程（Map）和归纳过程（Reduce）。MapReduce可以在一系列节点上执行并行计算，提升工作效率。

3 适合铁路客运的基于大数据平台的分析技术路线

大数据分析可分为4层，由低到高依次为：响应型分析、诊断型分析、战略分析和预测型分析；每层分析对应的产物为：报表，趋势分析，智能引擎与预测模型，行动建议。在目前铁路客运数据分析中，使用最多的仍然是响应型分析，分析人员每天为铁路总公司和铁路局的业务管理人员提供大量的统计报表；诊断型分析和战略分析也在使用，但使用的范围要比响应型分析小，预测型分析使用的范围更小。

结合铁路客运分析现状，基于大数据平台的铁路客运分析技术路线应该包括以下4部分：

（1）加强企业数据仓库建设；

（2）统筹利用铁路总公司和各铁路局分析平台资源；

（3）使用开源技术不断拓展铁路客运分析的数据源；

（4）建立企业分析数据集。

3.1 加强企业数据仓库建设

基于我国铁路客运发展的现状，统计报表在展现运输生产状况供领导决策过程中还发挥着重要的作用，因此还要加强企业数据仓库的建设，确保平台的高效性、数据的完整性、模型的多样性，在完成正常的统计报表的同时，能够满足更多的专项分析要求。

3.2 统筹利用铁路总公司和各铁路局分析平台资源

在大数据分析领域，云计算是非常重要的一项技术；通过合理地设置业务规则，使用MapReduce技术相关理念，将铁路总公司及各铁路局的分析平台组建成一个私有云平台，在保证数据安全性的同时可以极大地提高数据存储空间、提升运算效率。

3.3 使用开源技术不断拓展铁路客运分析数据源

大数据相关技术尤其是开源技术的不断进步实现了对之前无法想象的新数据源进行抽取、存储和分析，数据是最重要的价值驱动因素，特别是某一类新数据源的使用，对企业价值的挖掘具有非常重要的意义，这就是企业都在不遗余力地收集和使用各类可用大数据源的原因。通过对开源技术的使用，逐步实现对铁路客票系统所产生的半结构化、非结构化的数据进行分析、挖掘，并将其转化为数量不是很庞大的结构化数据存储在数据仓库中。

3.4 建立企业分析数据集

使用铁路客运大数据平台进行分析的用户有很多，包括铁路总公司、18个铁路局（公司）、中国铁道科学研究院等客票系统团队的分析人员。为了提升效率，使不同分析人员的分析成果具备可复制性，需要逐步建立企业分析数据集。企业分析数据集相当于大数据分析的一个中间层，包括每个分析人员常用的各种属性和分析指标，通过建立合理的管理流程，该中间层的属性和指标会逐渐丰富起来，分析人员可以共用这些属性和指标，更有利于从数据中分析挖掘出价值。

4 结束语

得数据者得天下，这里所说的数据不仅仅是作为资源保存在相关设备中的数据，更加重要的是指通过数据分析和挖掘，能够将蕴含在数据内的价值转化为促进铁路客运发展的生产力。铁路客运大数据分析平台处于起步阶段，在充分使用和借鉴当今流行的云计算、MPP、MapReduce等技术的基础上构建适合我国铁路客运发展的大数据分析平台，必将对我国铁路客运的发展和客运数据价值的挖掘起到重要的推动作用。

参考文献：

[1] Bill Franks. 驾驭大数据 [M]. 黄　海，译. 北京：人民邮电出版社，2013.

[2] B Lublinsky. Clearing the Air on Cloud Computing(April 22, 2009) [EB/OL].http://www.infoq.com/news/2009/04/air(retrieved June 03, 2009)

责任编辑　付　思