文章编号：1005-8451（2016）09-0017-05

铁路客运大数据平台的数据采集技术研究

阎志远，翁涅元，戴琳琳，杨立鹏
（中国铁道科学研究院 电子计算技术研究所，北京 100081）

摘 要：简要介绍大数据的概念、背景及其在铁路客运领域的应用意义，分析铁路客运大数据平台的数据源特点和需求特点，举例陈述现有大数据采集的相关技术以及软件应用，提出对铁路客运大数据平台数据采集方案的构想。

关键词：大数据；数据采集；铁路客运；客票系统
中图分类号：U293：TP39 **文献标识码：**A

Big data collection technology in railway passenger transport platform

YAN Zhiyuan, WENG Shengyuan, DAI Linlin, YANG Lipeng
(Institute of Computing Technologies, China Academy of Railway Sciences, Beijing 100081,China)

Abstract: This article summarized the concept and background of big data, and the significance of application of big data in the field of railway passenger transport, analyzed the characteristics of the data source and demand of the big data platform, described related techniques and software applications of existing big data collection, proposed the data collection scheme of big data platform for railway passenger transport.

Key words: big data; data collection; railway passenger transport; Ticketing and Reservation System

“大数据”来自于未来学家托夫勒所著的《第三次浪潮》，指无法在一定时间范围内使用常规软件工具进行捕捉、管理以及处理的数据集合。对于大数据的处理不使用随机分析法（即抽样调查法），而采用所有数据进行分析处理。大数据具有4V特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）[1]。

铁路客票发售与预订系统是覆盖全路的大型计算机广域网实时交易系统，实现了铁路客票的全国联网异地售票，计算机售票车站达2 400多个，约有12 000个计算机售票窗口投入运行[2]。2011年6月基于铁路客票发售与预订系统构建的12306互联网售票系统成功上线，尖峰日PV（Page Views）接近300亿，每秒出票超1 000张。随着客票系统的运行，每日将产生出海量的交易记录、旅客查询记录、页面点击行为等记录，这些记录数量庞大、产生速度极高、种类多样、隐含价值高、反映了旅客的真实特征与需求，符合大数据的特性，传统的数据处理手段已经不能满足庞大的数据处理需求。建立铁路客运大数据平台，对数据进行有效采集、存储并分析，将海量数据转化为价值是亟待解决的问题。

1 铁路大数据应用的意义

在铁路系统加快转变发展方式的新形势下，铁路大数据的应用可以帮助决策者从海量数据中挖掘旅客需求的内在规律，对市场做出更准确的分析与预测，从而做到快速把握市场变化规律、掌握旅客需求动向，及时制定有效的客运产品策略，有助于更好地构建铁路运输企业的核心竞争能力，提高铁路的盈利能力以及可持续发展能力。铁路大数据应用具有极其重要的现实意义[3]。

（1）提高信息整合能力。我国铁路系统组织机构庞大、分工复杂，信息的采集与流动具有分散性，不同业务、不同部门间容易出现信息割裂现象，导致铁路管理信息碎片化。大数据技术将若干分散数据源的数据进行集中采集，将异构数据整合在一起，有助于实现铁路管理信息的一体化管理。（2）建立市场导向的营销体系。大数据相关技术手段有助于利

收稿日期：2016-06-15
基金项目：铁路总公司科研计划课题（2016X005-A）；铁路总公司科研计划课题（2016X005-B）；铁路总公司科研计划课题（2016X004-G）；铁科院科研专项课题（研发中心）（J2016X009）。
作者简介：阎志远，副研究员；翁涅元，研究实习员。

用海量数据挖掘、获取客户、业务相关信息，提高中国铁路总公司营销部门的信息获取能力，为管理决策提供强大的数据支持。(3) 提高决策效率。通过对海量信息的实时处理，可以显著提高数据获取的时效性。相较于历史数据分析或市场调查，大数据技术可以帮助决策者及时掌握市场各方面数据信息，从而提高决策效率。(4) 对客运数据各业务处理流程中的数据产生、传输与处理过程进行采集与记录，有助于从各维度、各层次掌握客运业务的开展状况。

2 现有大数据采集关键技术与应用

大数据研究中涉及的数据具有来源丰富、分布广泛、格式多样、非结构化等特点，我们在对数据进行最终分析前，需要对这些数据进行采集、加工与存储。由于信息具有时效性，我们在对大数据进行有效处理的同时，也需要保证处理的高效性。

传统数据采集来源单一，所存储、管理的数据量相对较小，因此大多采用关系型数据库以及数据仓库进行处理。对于数据的并发处理而言，传统数据处理追求高一致性与容错性，而依据 CAP 理论 [4]，传统数据库无法保证可用性与扩展性。因此，传统数据处理方法已不能适应大数据模式的需求。

2.1 现有大数据采集关键技术

大数据的采集注重对数据的高速、实时处理，同时也需要大容量、高速、可扩展的存储系统来容纳数据。在大数据的采集与处理过程中所涉及的关键技术如下所述。

(1) 分布式存储

分布式存储系统将数据分散存储在多台独立的设备上，避免了传统存储的性能瓶颈，也提高了系统的可靠性和安全性。分布式存储系统中的 HDFS (Hadoop Distributed File System) 是应用最为广泛的系统之一。HDFS 是一个高度容错的系统，拥有故障检测和自动快速恢复特性，典型的 HDFS 文件大小是 GB 或者 TB 级别，支持千万级别数量的管理，适合大规模数据集应用。

一般 HDFS 存储系统以集群形式存在，为主从结构。HDFS 节点分为数据节点与名字节点，名字节点管理文件命名空间和调节客户端对文件的访问，数据节点负责数据的存储。文件在 HDFS 中被分割为多个块，分布在不同数据节点中，可实现数据的高速并发读取以及安全备份。名字节点作为 HDFS 的仲裁者和元数据仓库，代替用户管理数据节点的文件块，从而简化了用户的操作。

建立在 HDFS 之上的计算应用支持调用节点的计算能力，在数据集特别巨大的时候，将计算节点迁移到离数据更近的位置，而不是迁移数据到计算程序节点附近，从而消除了网络的拥堵，提高了系统的计算效率。

(2) 批量计算

批量计算数据是大数据计算应用场景之一，属于先存储后计算。对于实时性要求不高同时对数据的准确性、全面性更重视的应用领域中，批量计算模式更为适合。Hadoop 是典型的批量计算架构，由 HDFS 负责文件的静态存储，并通过 MapReduce 将计算逻辑分配到各数据节点中进行数据的计算以及价值发现。

(3) 流式计算

对于实时性要求严格，但数据精确度要求较为宽松的应用场景，流式计算具有明显优势。流式计算直接将流动的数据在内存中进行实时计算，数据往往是最近时间窗口内的，因此数据延迟往往较短，实时性强。

(4) 数据 ETL

数据抽取、转换和装载 (ETL，Extraction，Transformation and Loading) 是数据分析、获取高质量数据的关键环节。ETL 负责将异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后，进行清洗、转换、集成，最后加载到数据仓库或者数据集市中，成为联机分析处理 (OLAP，Online Analytical Processing)、数据挖掘的支持数据。在数据仓库构建的过程中，ETL 工作占整个工作的 50% ~ 70%，只有对数据进行了有效的处理，才能保证更好的 OLAP 质量。

2.2 现有大数据采集相关应用

铁路客运相关大数据一般源于各业务系统的日志数据，随着日志数量的不断增长，日志中所包含的信息也越显重要。目前，已有的日志数据采集应

用一般具有如下特征：应用系统与统计系统的解耦性、高度可扩展性、支持接近实时统计特性以及支持离线统计特性。现有日志采集应用列举如下。

（1）Flume

Flume 是 Cloudera 开发的高可用、高可靠的分布式海量日志采集、聚合和传输系统，该系统于 2011 年经重构后改名为 Apache Flume，Flume 系统架构如图 1 所示。

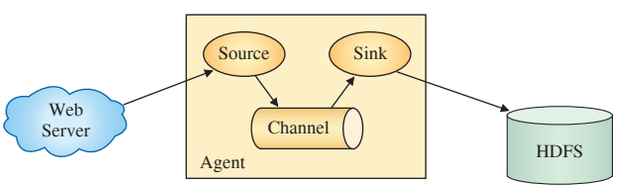

图1　Flume系统架构示意图

Flume 的主要概念包括：Client，部署在 Web Server 处，负责生产数据；Source，负责收集 Client 产生的数据并传递给 Channel；Channel，作为连接 Source 和 Sink 的通道，类似于队列；Sink，从 Channel 中获取数据并写入文件系统。Flume 提供了大量 Source、Channel 和 Sink 实现类型，不同类型均有各自的特点与功能。通过用户配置文件进行组合，Flume 可以组成极其灵活多变的日志收集系统 [5]。

（2）Kafka

Kafka 是 2010 年 12 月份开源的项目，采用 Scala 语言编写，使用了多种效率优化机制，整体架构（Push/Pull）比较新颖，更适合异构集群。

Kafka 实际上是一个消息发布订阅系统，主要有 3 种角色，分别为 Producer，Broker 和 Consumer。Producer 向某个订阅主题发布消息，而 Consumer 订阅某个主题的消息，一旦有某个主题的消息更新，Broker 会传递给订阅它的所有 Consumer。在 Kafka 中，消息是按主题组织的，而每个主题又会分为多个分区，这样便于管理数据和进行负载均衡。Kafka 系统架构如图 2 所示，其中，终端应用（Front End）即为消息的 Producer，Hadoop 集群、实时监控（Real-time monitoring）、其他服务、数据仓库则为消息的 Consumer 来接受消息。

（3）Spark

Spark 是加州大学伯克利分校的 AMP 实验室（UC Berkeley AMP lab）所开源的类 Hadoop MapReduce

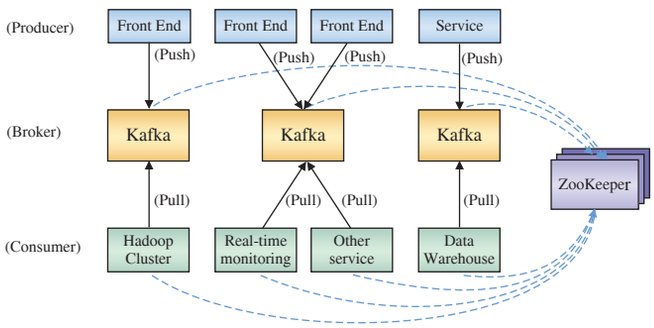

图2　Kafka系统架构示意图

的通用并行框架，Spark 的中间输出结果可以保存在内存中，从而不再需要读写 HDFS，因此，Spark 能更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 算法。

（4）Scribe

Scribe 是 Facebook 开源的日志收集系统，在 Facebook 内部已经得到大量的应用，其系统架构如图 3 所示。它通过 Scribe 代理（Scribe agent）从各种日志源上收集日志，存储到一个中央存储系统（网络存储系统 NFS 或者分布式存储系统 HDFS 等）上，以便于进行集中统计分析处理。它为日志的“分布式收集，统一处理”提供了一个可扩展、高容错的方案。它最重要的特点是容错性好。当后端的存储系统出现异常时，Scribe 会将数据写到本地磁盘上，当存储系统恢复正常后，Scribe 将日志重新加载到存储系统中。

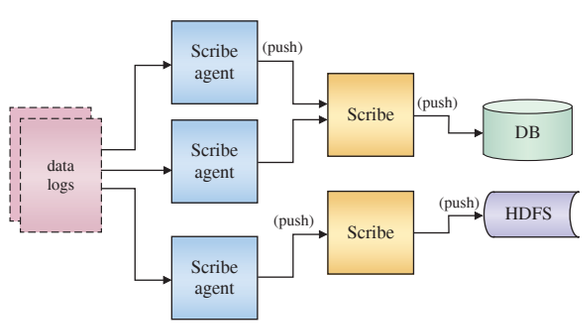

图 3　Spark系统架构示意图

3　铁路客运大数据平台数据源及需求特点分析

铁路客运相关业务拥有不同的需求及应用场景，因此各业务的数据源特点与应用要求不同，需要对各数据源进行分析。以下分别从售票数据、车站数据以及列车数据 3 个方面进行描述。

3.1　售票数据

客票销售包括电话售票、窗口售票、12306 网

站售票、手机 App 售票以及自动售票机售票等渠道。客票系统升级到 5.0 后，客票业务实现了核心交易数据的逻辑集中，所有客票应用都通过连接交易管理服务器（CTMS，Connection and Transaction Management Server）请求访问客票业务数据。因此，窗口、铁路局、铁路总公司等各级 CTMS 服务节点即可作为售票数据的数据源 [6]。

对售票数据的实时数据进行分析，可以及时了解旅客购票需求的热点信息，根据需求热点指定灵活的票额分配以及票价浮动策略，以提高客票销售的收益水平。

3.2 车站数据

旅客进站前的身份证以及车票信息核验数据为实名制数据，因此，对进站安检信息进行采集与分析可以较为准确地掌握各车站实时进站客流数量，结合当日各客票销售数据所统计的车站发送客流量信息，可以及时发现车站客流数量的异常，达到安保预警的目的。

旅客可以使用车票、身份证通过站台闸机，通过对闸机验票数据的采集与分析，可以从中提取各站台实时客流数量以及旅客的换票行为特征等信息，以上信息有助于更好地进行车站客运组织。

3.3 列车数据

通过列车员手持设备可获取车上旅客的实名信息及车票信息，通过信息核验即可得知旅客是否具有乘车资格，标记无乘车资格的乘客信息并及时上报，待旅客下车后，车站工作人员即可根据车上设备采集的数据及时处理。

综上所述，旅客在购票、进站、乘车、出站各环节中产生的数据均可作为大数据平台的数据源。通过对数据的采集与整合，即可掌握旅客乘车出行的全部环节的行为特征。

4 铁路客运大数据平台数据采集方法构想

大数据的采集涉及数据的产生、处理和存储 3 个阶段。基于铁路客运的数据源及需求特性，铁路客运大数据平台的数据采集需要同时兼容传统数据源以及互联网售票等新型业务的非结构化数据。以下分别从平台结构设计、数据源组织等方面进行说明。

4.1 平台结构设计

为最大程度利用既有项目投资，铁路客运大数据平台对于客运业务数据的采集无需从车站、铁路局开始，是从 CTMS 连接交易管理器中获取交易数据。而对于互联网业务等日志数据的采集与处理，则通过结合目前开源的日志采集技术、消息队列技术、数据流处理技术以及分布式存储技术，实现大量数据的实时收集、传输、处理与存储。

平台结构设计如图 4 所示。

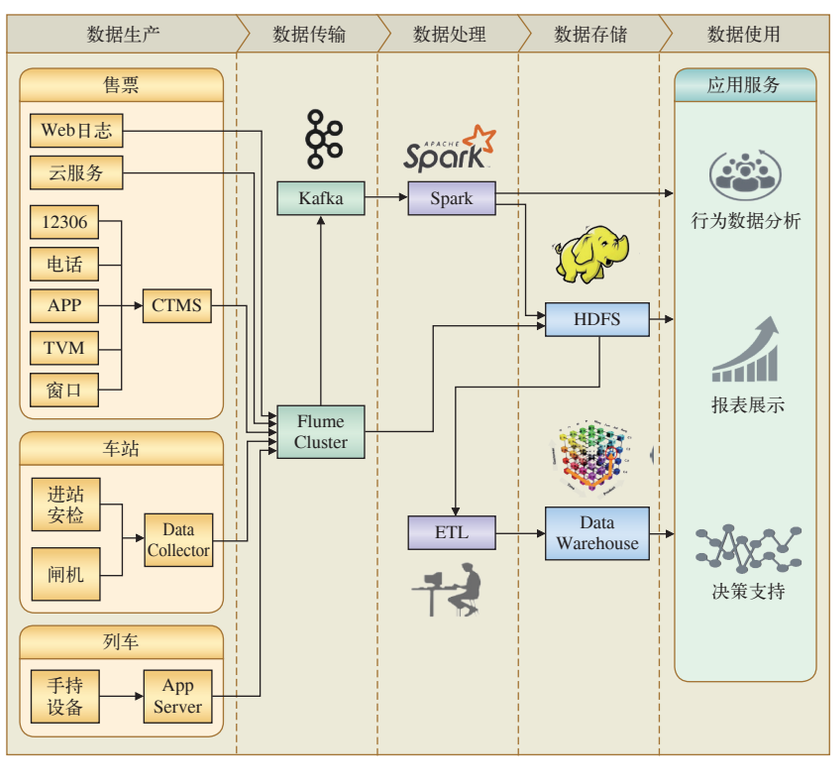

图4 铁路客运大数据平台结构设计

对于 12306 网站、电话、手机购票 App、自动售票机以及窗口的售票数据均由 CTMS 转发至 Flume 集群，网站 Web 服务以及云服务日志数据则直接由 Flume 集群收集。

由于车站安检设备以及闸机配置多样，因此需要在车站设置数据采集器，由数据采集器统一收集，并对日志进行规范处理后转发至 Flume，完成车站数据的采集。

列车车载设备由于受网络通信条件以及硬件设备的限制，不能保证数据的实时传输，因此，可以

把位于地面的车载设备服务器视为数据源，将各车载设备的交互信息通过大数据平台进行采集。

Flume 集群在接受到日志信息后，将数据存储在分布式存储系统中，同时将需要实时分析的数据通过 Kafka 传递至 Spark 进行流式处理，并将处理结果存储在分布式存储中。已保存在分布式存储的历史数据通过 ETL 处理转化后，保存在数据仓库中，供数据使用者进行进一步分析。

在数据使用层部署数据查询应用服务器以及分布式计算服务器集群，为数据使用者提供方便的数据查询和使用服务接口。用户可以直接调用接口查询数据或者利用计算集群的计算能力进行数据挖掘与分析工作。

4.2 数据存储设计

由于铁路客运大数据平台服务应用需求的不同，对于不同性质、不同时期的数据应该以不同方式进行处理。

（1）运营数据

客运营销分析和数据挖掘，需要查询与分析大量历史数据，但是对于查询时间要求不高。此类数据可以存储在传统数据仓库或者分布式存储中。对于存储于传统数据仓库中的关系型数据，可以进行索引优化以加快数据的查询速度；对于分布式存储数据的使用，则可以结合分布式计算技术，加快数据的查询与处理速度。

（2）热点数据

基础字典、票价信息、余票信息等需要被频繁查询的热点数据要求极低的查询延时，可以将此类数据存储在内存数据库中，以加快查询速度。

（3）备份数据

大数据平台需要保障系统运行期间积累下来的历史数据的安全。通过对数据进行冗余备份可以极大地减少数据损害丢失的风险，同时对于此类数据可以通过压缩存储的方式减少数据存储的成本。

5 结束语

本文对大数据的应用与意义进行了概述，分析了铁路大数据平台的数据特点与需求，对现有大数据采集技术与应用进行了陈述，并对铁路客运大数据平台的数据采集系统架构进行了构想。

大数据的应用是铁路将数据转化为竞争力的必然选择，对于提高铁路信息化水平、构建市场导向的营销体系、提高信息整合能力、提高决策效率具有重要意义。数据采集作为大数据应用的基础与必要步骤，其意义更是重中之重。

参考文献：

[1] 孟小峰，慈　祥．大数据管理：概念、技术与挑战 [J]. 计算机研究与发展，2013, 50（1）：146-169.

[2] 朱建生，单杏花，周亮瑾，等．中国铁路客票发售和预订系统 5.0 版的研究与实现 [J]. 中国铁道科学，2006, 27（6）：95-103.

[3] 代明睿，朱克非，郑平标．我国铁路应用大数据技术的思考 [J]. 铁道运输与经济，2014, 36（3）：23-26.

[4] 陈　明．分布系统设计的 CAP 理论 [J]. 计算机教育，2013（15）：109-112.

[5] Chambers C, Raniwala A, Perry F, et al. FlumeJava: easy, efficient data-parallel pipelines[J]. AcmSigplan Notices, 2010, 45(6): 363-375.

[6] 阎志远，王智为，张常顺，等．铁路客票 CTMS 架构和关键技术 [J]. 铁路技术创新，2012（4）：26-28.

责任编辑　付　思