文章编号：1005-8451（2016）06-0018-06

开源商业智能技术在铁路客运营销辅助决策中的应用

汪健雄，王　芳，贾成强，刘婷婷
（中国铁道科学研究院　电子计算技术研究所，北京　100081）

摘　要：提出基于Pentaho的铁路客运营销辅助决策解决方案，重点介绍基于Kettle的ETL设计、报表立方体设计和OLAP分析、报表制作以及根据用户需要定制个性化报表控件等关键技术，解决了商用软件由于在接口和代码上的封闭性，不能完全自定义客户个性化需求的问题，节约了软件采购成本，具有广阔应用前景。但该解决方案也存在一些不足，如：开发工作量大、软件文档不足，以及系统安全性和并发性能未做很好的测试和优化等问题，在今后的研究和生产中需要重点解决。

关键词：开源商业智能；Pentaho；客运营销辅助决策；OLAP；自定义控件
中图分类号：U239：F530.86：TP39　**文献标识码：**A

Open source business intelligence technology applied to Railway Passenger Marketing aided Decision System

WANG Jianxiong, WANG Fang, JIA Chengqiang, LIU Tingting
(Institute of Computing Technologies, China Academy of Railway Sciences, Beijing　100081, China)

Abstract: This article proposed a solution of Railway Passenger Marketing aided Decision System, introduced several key technologies, such as Kettle based ETL design, report cube design, OLAP analysis, report making and custom personalized report control, and so on. This solution could overcome the shortage of closing codes and hardness of custom control, save software cost, and lead to a wide used prospect. Otherwise, some problems should be settled in the future, such as a large amount of coding work, lack of documents, insufficient safety test and parallel performance test.

Key words: open source business intelligence; Pentaho; Railway Passenger Marketing aided Decision System; OLAP; custom control

铁路客运营销辅助决策系统是为铁路总公司、各铁路局提供铁路运能、运量、收入、效益分析等指标的决策支持系统。从系统建设之初起，引入商业智能软件，实现了客运数据在铁路总公司和铁路局范围内的共享，铁路总公司、铁路局、站段各级用户通过统一授权访问，从运能、运量和收入各个层面分析、评价客运组织情况，预测客流趋势并指导今后的发展，铁路各级管理者在客运组织工作方面有了重大改变，起到了提高客运业务的核心竞争能力的作用[1]。但上述商业智能软件在适应了复杂业务分析的需求之后，一些问题逐渐显现出来，主要包括：

（1）商用软件在接口、代码上的封闭性，不能完全自定义客户个性化的需求，如果厂商不主动发布升级版本，很难对系统进行拓展。

（2）随着数据量与日俱增，当前商业智能软件在大数据应用方面的支持相对较弱，同时系统查询性能也存在瓶颈，很难再继续优化。

（3）商用软件价格昂贵，支出较大，不利于系统建设方控制成本。

近年来，开源商业智能项目在互联网行业得到了长足发展，很多优秀的开源产品可与商业产品一较高低。在商业智能（BI）方面，近年来开源社区中的Pentaho具有很多特点，正成为开源BI事实上的标准，为此在铁路总公司级的客运营销辅助决策系统中尝试采用开源的Pentaho BI套件解决方案。

1　Pentaho BI套件简介

Pentaho是对多个开源项目进行改进、扩充和集

收稿日期：2015-12-03
基金项目：中国铁道科学研究院基金项目（2014YJ013）。
作者简介：汪健雄，副研究员；王　芳，副研究员。

成组成的 BI 平台，涵盖了常规 BI 系统的开发、部署和运行环境。Pentaho 平台的核心思想是以业务流程为核心，基于工作流技术，让决策成为业务的一个环节，实现企业业务过程整合 [2]。Pentaho 提供了围绕特定项目制定方案的集成开发环境，体现了面向解决方案的 BI 研发思路。PentahoBI 平台主要由以下几部分组成：

（1）OLAP 服务器：集成了基于 Java 开发的 OLAP 服务器，Mondrian，用于对存储在关系数据库中的大型数据集进行交互分析。

（2）OLAP 分析工具：集成了 JPivot 可视化组件，可实现多维数据表和多维数据图以及数据立方体的展示。报表工具组件名为 ReportDesigner，是基于 JSP 的 B/S 分析工具，用于自定义分析报表。

（3）ETL 组件 PDI：Pentaho 整合了开源 ETL 工具 Kettle，包括 Spoon 和 Pan 两个包。Kettle 提供的 Spoon 和 Chef 工具提供 Drag&Drop 的图形化界面，用于定义和执行 ETL 转换流程，同时在 Chef 或 Kitchen 中通过 Job 可以定义和执行定时任务。

（4）数据挖掘工具 Weka：Weka 作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。

（5）集成管理和开发环境：Pentaho Design Studio 是基于 Eclipse 的开发、项目测试和部署环境，集成 Action Sequence 编辑器用于定义工作流的图形化界面。

Pentaho 涵盖了数据仓库、ETL、OLAP、数据挖掘以及报表生成、仪表盘等应用的测试和部署的集成开发环境，是目前对 BI 的功能支持最为全面的开源套件，同时与商业软件相比在二次开发和成本方面具有较大优势，因此选用 Pentaho 进一步研究铁路客运决策支持系统。

2 基于Pentaho的铁路客运营销辅助决策系统设计

基本设计思路是以数据仓库的设计和实施为中心，数据挖掘的应用为补充，构建基于 Pentaho 的商业智能系统。通过在铁路总公司营销系统及其他业务系统基础上构建 Pentaho 商业智能平台来实现系统集成，使从日常的业务中的操作型数据变为分析型数据，从分析型数据中提炼决策信息，协助铁路客运管理者做出正确的决策。

系统分为 4 个层次：

（1）数据层：包括原有的中国铁路总公司营销系统、客票发售与预定系统（简称：客票系统）可以为 BI 提供大量的宝贵的源数据，同时为了解决铁路总公司综合分析的要求，引入了客图接口数据、铁路客运清算系统成本数据以及其他可用于客运决策支持的原始数据。

（2）基础架构层：引入 Pentaho BI 平台中的 OLAP 技术和 Weka 数据挖掘工具进行多目标、多维度的分析以及即席查询；未来还将引入开源数学计算项目 R 来实现预测、盈亏分析等应用的模型与算法形成运算引擎。通过基于 Kettle 的数据抽取、转换、加载工具形成数据仓库。

（3）业务应用平台：根据铁路总公司需求，重点实现可图管理、数据挖掘、运营报表、预测和盈亏分析等应用，该平台将集成在 Pentaho BI Server 组件中。

（4）展现层：根据铁路总公司需求，利用 JSP、AJAX、Flex 等技术实现报表、统计图形、OLAP 展现以及一些自定义交互应用。其中，仪表盘可以高效集成各种 BI 内容，并以较简单、统一的视图呈现给各级用户，各种不同层次的 BI 用户还可以定制适合自己的仪表盘。Pentaho Dashboard 工具基于 CDF（Community Dashboard Framework）项目整合而来的，可以直接将仪表盘等应用集成在 Pentaho BI Server 中作为解决方案进行发布，如图 1 所示。

3 Pentaho BI技术在铁路客运营销辅助决策中的应用

3.1 ETL设计

数据的抽取、转换和加载 (ETL) 是 BI 项目中最常见、基础的数据操作。在数据仓库的构建期间，各个业务系统的数据必须经过严格的 ETL 过程，整合到数据仓库中为后续的分析、数据展现提供支撑。数

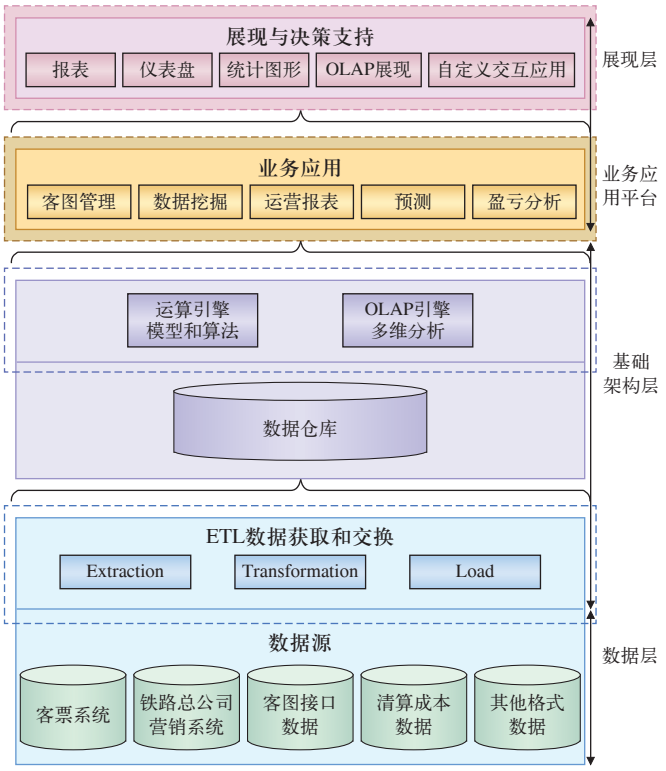

图1 基于Pentaho的客运营销BI体系设计

据可能来自不同业务部门，以及不同的数据源规格。另外，一些即席报表对运行的时间要求较高，通常需要对海量数据进行数据聚合和初步加工来更改数据的粒度，使得报表服务器可以更快的响应用户提交的数据请求。在 Pentaho 平台使用 Kettle 作为 ETL 处理组件，从 SybaseASE 数据源、SybaseIQ 和平面文件中抽取数据。利用 Kettle 中的 Spoon 工具对业务数据进行必要的字段处理和格式转换，把处理过的数据重新加载到数据仓库中，然后利用 Kitchen 工具，实现系统定期执行 ETL 脚本，完成数据的自动抽取 [3]。Kettle 工具定义的 ETL 流程如图 2 所示。

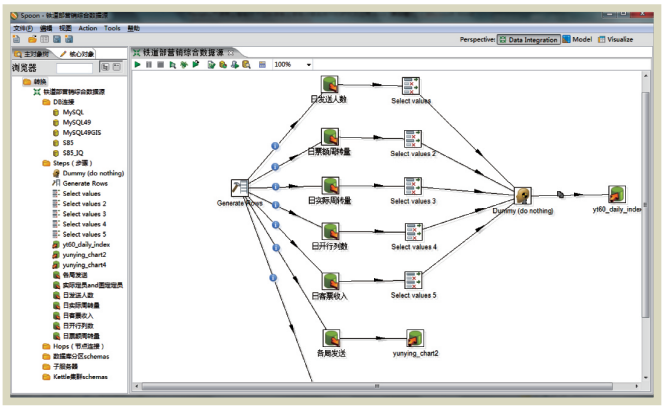

图2 Kettle工具定义的ETL流程

3.2 OLAP分析

Pentaho 平台使用的 Mondrian 组件提供 OLAP 服务。利用可视化工具 Pentaho Schema Workbench，分析人员可以编写多维查询语言 MDX，并形成模板提供给分析人员使用，同时对数据仓库中的数据集进行交互分析 [4]。为便于试验，使用 Tomcat 作为 Mondrian 应用服务器。使用 Mondrian 的架构进行 OLAP 分析的具体步骤如下：

（1）底层：数据库或从原有的数据库中提取可用的数据表。

（2）存储层：数据仓库的建立，将原有的底层数据库转化为星型模型或雪花模型。

（3）维度层：生成 schema 文件，将存储层的数据仓库转化为一个 schema 文件，通过 schema-workbench 或者手写完成，可以通过 MDX 来对多维数据库进行访问，并产生可部署到 Mondrian 服务器上的 OLAP 解决方案，解决方案的基本配置文件包括流程文件 xaction、立方体描述文件等。Mondrian OLAP 引擎根据部署的配置文件，从数据库中计算和缓存数据，并响应来自展示层的各种查询。专业分析用户可以直接使用 MDX 语句访问；将 MDX 预先存储后，非专业用户也可以在图形化交互界面中进行数据分析。

（4）展示层：编写 jsp 文件用于 OLAP 展示，由 JPivot 提供的表现层 TagLib 实现，这是一个使用 Web 组件框架 (WCF) 技术、采用 XML/XSLT 渲染 Web UI 的开源组件，可以比较方便的将多维数据展示给最终用户，可展现多维数据透视图表，支持钻取、切片、旋转等操作。一个按运营单位、日期、票种 3 个维度分析发送量的立方体模型的 schema 定义如下：

```
<Schema name="SendAnalysisSchema">
<!-- 日期维度 -->
<Dimension type="StandardDimension" visible="true" name="TrainDate">
<Hierarchy allMemberName="AllDate" primaryKey="train_date">
<Table name="dic_date" alias=""></Table>
<Level name="Train Date" table="dic_date" column="train_date"></Level>
```

```
</Hierarchy>
</Dimension>
<!-- 车站维度 -->
<Dimension type="StandardDimension" name="Station">
<Hierarchy allMemberName="AllStation" primaryKey="station_telecode">
<Table name="dic_station" alias=""></Table>
<Level name="Send Station" visible="true" table="dic_station" column="station_telecode"></Level>
</Hierarchy>
</Dimension>
<!-- 票种（客流类型）维度 -->
<Dimension type="StandardDimension" visible="true" name="TicketType">
<Hierarchy allMemberName="AllTicketType" primaryKey="ticket_type_code">
<Table name="dic_ticket_type" alias=""></Table>
<Level name="Send Type" visible="true" table="dic_ticket_type" column="ticket_type_code">
</Level>
</Hierarchy>
</Dimension>
<Cube name="Send Analysis" visible="true"cache="true" enabled="true">
<Table name="F_send_work" alias="">
</Table>
<DimensionUsage source="Train Date" name="Access Time"visible="true" foreignKey="fTime">
...</DimensionUsage>
<!-- 人数指标 -->
<measure name="up_num" column="um_num" aggregator="sum" visible="true"></Measure>
<!-- 收入指标 -->
<measure name="income" column="income" aggregator="sum" visible="true"></Measure>
</Cube>
</Schema>
```

Cube 建好后即可以利用 Kettle 抽取数据，并使用 JPivot 生成报表。

3.3 自定义报表

Pentaho 提供的报表生成工具为 Report Designer，可以根据用户需要制作专业化的分析报表，并支持 Excel 或 PDF 等通用格式的展现[5]。图 3 所示的是正在编辑报表的界面：左侧的竖排工具栏显示的是设计报表时可能用到的控件。中间的部分是编辑自定义报表的主界面，右边的标签 Structure 可以看到报表各个元素的结构，Data 标签包含了展示的数据，如包括报表 query 的数据源及各种函数。报表设计主界面分成了 PageHeader、ReportHeader、Details、ReportFooter、PageFooter 等多个区域。Page Header 与 Page Footer 中的对象会在报表的每页都显示。Report Header 中的对象只在报表开头时展示一次，Detail 中的对象会展现 query 中的结果集，Report Footer 中的对象只在报表的末尾显示一次。需要展示的字段放在 Details 区域，通过 Structure 标签可以查看报表数据项与页面之间的组成关系。

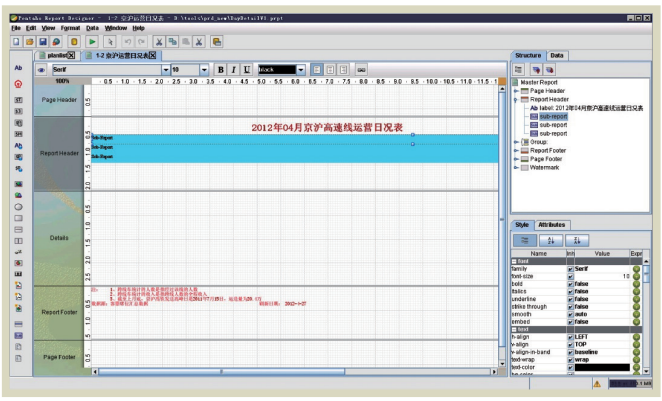

图3 使用Report Designer实现的报表设计视图

Report Designer 制作的报表可以设置输入参数，通过利用参数来对报表数据源设置 filter 以达到传递交互式查询条件的目的。可在报表查询的主 query 中加入参数 StartDate（其格式为 ${StartDate}）。报表执行时将会显示当满足“train_date=${StartDate}”的值时，查询语句所选择的数据，如图 4 所示的京沪高铁运营日况报表。该报表除了 StartDate 参数外，提示页上还可以使用下拉框选择报表运行后输出的格式。输入参数以后运行的结果页面：

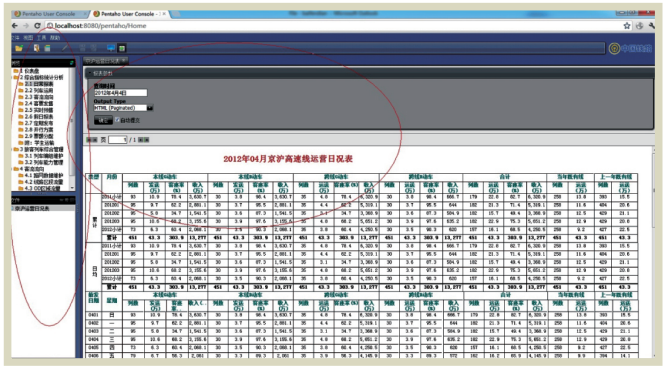

图4 在完成的BI门户中实现的报表运行视图

3.4 自定义报表控件

开源BI的优势之一是可以自定义报表控件，对开发工具进行扩展。在列车编组信息管理的开发中，需要实现分席别、指定区域，列车等级和方向列车图定与实际能力查询得到旅客列车对数表和运能统计表，同时需要在集成的图形中展示列车编组、交路信息，包括列车基本信息、运行里程、车底组数、编组布局、交路图形、开行规律、种类型号、运用状态等。现有使用的BI平台无法实现我们需要的效果。在Pentaho中，使用了JFreeChart这个开源工程来实现自定义报表控件[6]，并通过控件可视化拖拽的方式在其他报表中进行复用。

在meta-elements.xml中定义元件的元素定义：

```
<meta-data
xmlns="http://reporting.pentaho.org/namespaces/engine/classic/metadata/1.0">
    <include-globals
  src="res://org/pentaho/repor-ting/engine/classic/core/metadata/global-meta-elem-ents.xml" />
    <element name="station" hidden="false" bundle-name="metadata"
            implementation="StationChartType">

            <attribute-group-ref ref="interactivity" />

            <attribute namespace="http://reporting.pen-taho.org/namespaces/pr4jd"
                    name="bianzu"
                    mandatory="true"
                    hidden="false"
                    value-type="java.lang.String"
                    value-role="Value"/>
            ...// 一系列属性定义，如始发站、终到站、编组、对数等。
            </element>
    </meta-data>
```

在StationChartType.java文件中实现下列控件操作：

```
// This ElementType implementation renders a Star in a report
public class StationChartType implements ElementType {
        // 控件属性
        private String bianzu;// 编组信息
        private float duishu; // 对数信息
        // 在 meta-elements.xml 文件中定义的属性
        private transient ElementMetaData element-Type;

        // 默认构造函数
        public StationChartType() {     }

        // load the default metadata about the star element type
        public ElementMetaData getMetaData() {
                /* 获取元数据信息 */
        }

        public Object getValue(final ExpressionRun-time runtime,final Element element){
                /* 获取报表界面定制数值 */

        }

        XYDataset createDataset() throws ParseExce-ption {
                /* 获取数据源 */
        }
```

```
JFreeChart createChart(XYDataset param-
XYDataset) throws ParseException {
    JFreeChart localJFreeChart = Chart-Factory.
createTimeSeriesChart(
        null, " 时间 ", null, paramXYDataset,
false, false,false);
        /* 创建图表 */
        return localJFreeChart;
    }
    // 控件属性的 Setter 函数
    void setBianzu (String newBianZu) {this.
bianzu = newBianZu;}
    void setDuishu (float newDuishu) {this.duishu
= newDuishu;}
}
```

将整个文件包编译成 jar 包后放到 PRD 的 lib 目录下，重新启动 PRD 会在左侧控件栏中看到一个红色火车头为图表的自定义控件 StationChart，并可以将它拖放到报表中。

StationChart 属性和布局如图 5 所示，StationChart 的运行图报表如图 6 所示。

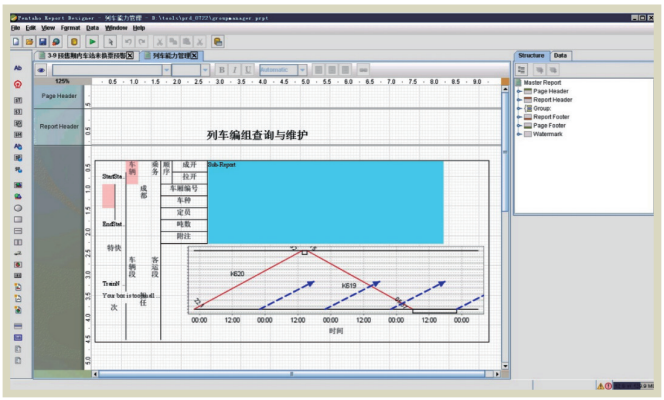

图5　在Report Designer中设置StationChart属性和布局

4　结束语

本文提出了在铁路客运营销系统基础上构建基于 Pentaho 的商业智能系统的方法，提出了基于 Kettle 的 ETL 设计、报表立方体设计和 OLAP 分析、报表制作以及根据用户需要定制个性化报表控件等解决方案，在一定程度上解决了由于商用软件在接口、代码上的封闭性，不能完全自定义客户个性化需求的问题，同时节约了软件采购成本。Pentaho 平台整合了一套适用于开发、部署和运行商业智能系统的环境与工具，为企业级商业智能实现提供了一种开放、经济的平台，进一步深入研究还有仪表盘、与 Hadoop 整合实现大数据集群等应用，由于篇幅问题将另作讨论。

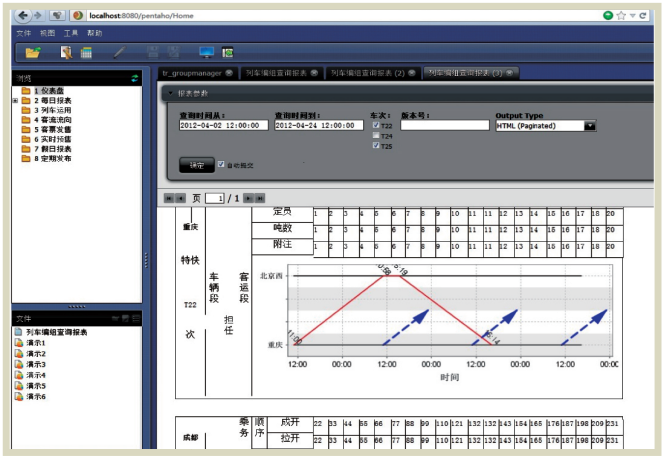

图6　在完成的BI门户中查看StationChart的报表运行视图

参考文献：

[1] 汪健雄，刘春煌，单杏花，等．业务智能技术在铁路客运营销辅助决策系统中的应用 [J]. 铁路计算机应用，2009，18（12）：23-27.

[2] Pentaho Corporation．Pentaho open source business intelligence platform technical white paper[EB/OL]．http://www.pentaho.com，2015.

[3] Pentaho Documentation Team. Evaluate and Learn Pentaho Data Integration[EB/OL]. https://help.pentaho.com/Documentation/5.3/0D0/1A0/010/000, 2015.

[4] 陈荣鑫，付永钢，陈维斌．基于 Pentaho 的商业智能系统 [J]. 计算机工程与设计，2008，29（9）：2407-2409.

[5] jfree.org．JFreeChart API Documentation[EB/OL]. http://www.jfree.org/jfreechart/api/javadoc/ index.html，2015.

[6] Pentaho Documentation Team. Business Analytics Report Designer[EB/OL]. http://www.pentaho.com/training-course/business-analytics-report-designer，2015.

责任编辑　徐侃春