

# **An Open-Source Solution of Business Intelligence in China Railway Passenger Transportation Decision Support**

Jian-xiong Wang<sup>1,a\*</sup>, Nan Wang<sup>1</sup>

<sup>1</sup>Institute of Computing Technologies, China Academy of Railway Sciences, 100081 Beijing, China

<sup>a</sup>wjx304@163.com

**Key words:** Open-Source; business intelligence; Decision Support; railway passenger transport marketing; Pentaho

**Abstract.** An Open-Source solution based on Pentaho is described in this paper for constructing railway passenger transport marketing assistant decision system as alternative of commercial software, solving the problem of being lack of user custom properties, which will save project costs, and has good expansibility. Furthermore, some key technologies are introduced including ETL design of kettle, cube design and OLAP analysis, reporting and self-customized report plugin, which may lead a deeper research.

## **1 Introduction**

Railway passenger transport marketing Aided decision system provides transportation capacity, traffic, revenue, benefit analysis and other indicators of China railway. Commercial software such as Business Objects, Cognos and other business intelligence suppliments are introduced, realized the sharing of passenger transport data within the scope of all levels of the China Railway, studying transport capacity, traffic and revenue as the index of evaluating the organization of passenger transport, which forecasts passenger flow trends and guides future development. Recent years, China Railway has got a significant change in passengers' management, an played a role in improving the core competitiveness of the business of passenger transport [1]. However, some problems also have gradually emerged, major problems include:

- Commercial software interfaces and closed of code cannot be completely adapted to the customer's personalized demand, if manufacturers do not take the initiative to release an upgraded version, it is difficult for the system to expand.
- For increasing amount of data, the current commercial business intelligence software is relatively weak to support big data applications, while the system query performance is also a bottleneck, it is difficult to continue to optimize.
- Commercial business intelligence software is expensive, large expenditure, an not conducive to the control of system construction costs.

Recent years, open-source projects have made considerable progresses in Internet industry, a lot of good open-source products are comparable with commercial products. In business intelligence, the open-source community's Pentaho has many features, is becoming a fact standard of open-source business intelligence, so we try to use Pentaho business intelligence Suite Solution in China passengers railway marketing assistant decision support system.

## **2 Pentaho Business Intelligence Suite**

Pentaho is a business intelligence platform based on improving, expanding and integrating some open-source projects, which covered business intelligence system development, deployment, and runtime environments. Pentaho platform's key idea is the core business process, based on workflow technology, make decision to be a part of the business to integrating enterprise business processes [2]. Pentaho provides integrated development environments around a specific project development plan, reflecting the business intelligence solution-oriented research and development ideas. The Pentaho's business intelligence platform consists of the following components:

- OLAP server: Integrating Java-based integrated development of OLAP server: Mondrian, interactive analysis of large data sets stored in a relational-database.
- OLAP analysis tools: JPivot, integrated visual components, can be realized the multi-dimensional data table, data diagram and data cube display. ReportDesigner is a Reporting tool component based on the B/S JSP analysis tool for custom analysis reports.
- ETL components PDI: Pentaho integrates the open-source ETL tool Kettle. Including Spoon and Pan two packages. Kettle provides Spoon and Chef tools to provide Drag&Drop graphical interface for the definition and implementation of the ETL conversion process, while in Chef or Kitchen through the Job can define and execute the timing task.
- Weka, data mining tool: Weka, Waikato Environment for Knowledge Analysis, works as a public data mining workbench, a large amount of data mining machine learning algorithms is built-in, including data pre-processing, classification, regression, clustering, association rules, and in the new interactive interface visualization.
- Integrated management and development environments: Pentaho Design Studio, based on the development of Eclipse, project testing and deployment environments, integrated Action Sequence editor for the definition of workflow graphical interface.

### **3 Railway passenger transport Marketing Decision Support System Design based on Pentaho**

The basic design idea is centered on the design and implementation of data warehouse, the application of data mining as a supplement, to build a business intelligence system based on Pentaho. Through constructing Pentaho business intelligence platform in the China Railway marketing system and other business systems to realize system integration, from daily business operational data are transferred to analysis data, for extracting decision information, which assists the management of railway passenger transport to make the right decision.

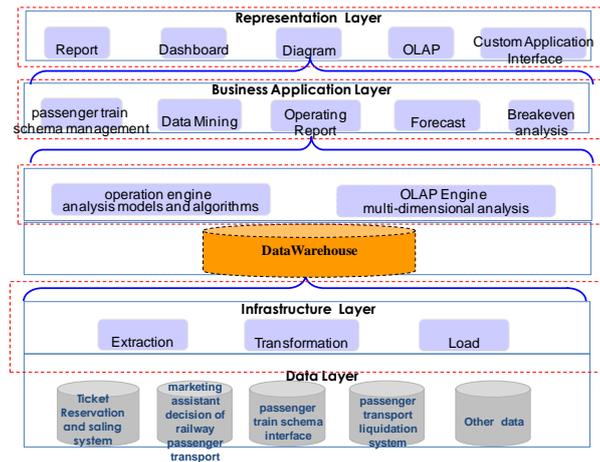
System is divided into 4 levels: data layer, basic infrastructure layer, application service layer and decision support layer.

- Data layer: China Railway marketing system and ticketing reservation and selling system can provide a lot of valuable data source for the business intelligence. In order to solve the comprehensive analysis, passenger train's interface data, the railway passenger transport liquidation system and other cost data, and other relative raw data are also introduced.

- Infrastructure layer: Introducing Pentaho business intelligence platform OLAP technology and Weka data mining tool to process multi-target, multi-dimensional analysis and ad-hoc queries; In the future, other open-source projects like R math library can be used for achieving forecast, profit and loss analysis models and algorithms of formation of operation engine. Through data extraction, transferring and loading tools to build data warehouse.

- Business application layer: According to China railway needs, Business application mainly focused on the implementation of the passenger train running information data management, data mining, operation reports, forecasts and profit and loss analysis and other applications, the platform will be integrated in the Pentaho business intelligence Server components.

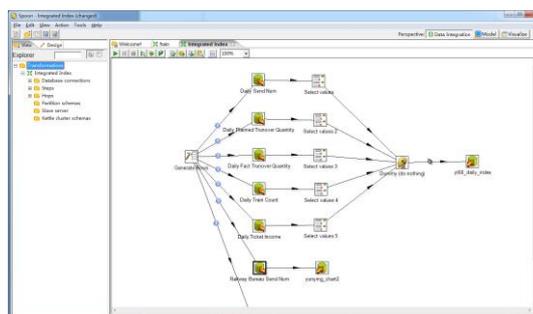
- Presentation layer: Some custom interactive applications are built on JSP, AJAX, Flex and other technologies to achieve reporting, statistical graphics, OLAP display. the Instrument panel (Dashboard) can be highly efficient integration of all kinds of business intelligence content, and with a simple, unified view to all levels of users, different users can also customize their own instrument panels. Pentaho Dashboard tool integration based on CDF (Community Dashboard Framework) from the project, can be directly integrated into the Instrument panel applications Pentaho business intelligence server as a solution for publication.



**Figure 1.** Design of business intelligence system for passenger transport marketing based on Pentaho. The following are several technical points involved in system development.

#### 4 key technologies

**4.1 Design of ETL.** Data extraction, transformation and loading (ETL) is the most common basic data manipulation in business intelligence project. During building a data warehouse, data of each business system must undergo a rigorous ETL process, integrated into the data warehouse for subsequent analysis and decision support. Data may come from different business sectors with different data sources specifications. In addition, some ad-hoc reporting on the running time-critical, often require massive data aggregation and data processing to change the Clustering of data, for making a faster corresponding from server after user submitting request. Pentaho platform uses Kettle as the ETL processing component, extracts data from Sybase ASE data source, SybaseIQ, and other flat files. The use of Kettle Spoon tool for business data necessary field processing and format conversion, the processed data is re-loaded into the data warehouse, and then Kitchen tools to execute ETL scripts to extracting data automatically [3].



**Figure 2** Kettle tools defined ETL process

**4.2 OLAP.** Pentaho platform uses Mondrian components to provide OLAP service. Using visualization tool Pentaho Schema Workbench, analysts can write a multidimensional expressions (mdx), and form templates. At the meanwhile, interaction analysis can be made through data set in data warehouse [4]. In order to facilitate the development, we use Tomcat as the Mondrian application server. The specific steps of using the Mondrian architecture for OLAP analysis are as follows:

- Design the underlying database or extract the available data tables from the original database;
- Storage layer: Establish data warehouse, transform the original underlying database into a star or snowflake model;
- Dimension layer: Generated schema file, transform storage tier data warehouse into a schema file by schema-workbench or by handwriting, and can be performed by MDX multidimensional database access, and produces deployed to OLAP solutions on the Mondrian server solutions, the basic profile includes process documentation x action, cube description files. Mondrian OLAP engine based on the profile deployment, calculated from the database and cache data, and in

response to various queries from the representation layer. Professional analysis users can access using direct MDX statements; through pre-storing MDX, non-professional users can analyze data in a graphical interactive interface.

• representation layer : Write JSP file for OLAP display , Performance layer (Tag Lib) implementation provided by JPivot, It's an open-source components based on Web component framework (WCF) technology, using XML/XSLT rendering Web UI, it can be convenient for multidimensional data presented to the end user, multi-dimensional data show perspective diagrams, supports drill, slice, rotation and other operations. A three dimensional cube model schema for analyzing the passenger transport volume, including operating unit, date and ticket kind dimensions is defined as follows:

```

<Schema name="SendAnalysisSchema">
  <Dimension type="StandardDimension" name="TrainDate"/>
  <Dimension type="StandardDimension" name="Station"/>
  <Dimension type="StandardDimension" name="TicketType"/>
  <Cube name="Send Analysis" cache="true" >
    <measure name="up_num" datatype="Integer" aggregator="sum" />
    <measure name="income" datatype="Numeric" aggregator="sum" />
  </Cube>
</Schema>

```

Figure 3 Cube Definition

Cube after the completion of the use of Kettle to extract data, and use JPivot to generate reporting.

**4.3 Reporting.** Pentaho provides a report generation tool (Report Designer), can create professional analysis reports according to user needs, and supports Excel or PDF and other common formats show[5]. The following figure 3 shows the interface to edit the reporting: reporting design controls are displayed on the left side of the vertical toolbar. the middle part is the main editing area of reporting. Structure tab on the right side, displays various elements of the reporting, Data tab show data, including the reporting query's a data source and a variety of functions. Reporting design main interface is divided into Page Header, Report Header, Details, Report Footer, Page Footer and other regions. The Page Header and Page Footer of the reporting objects on each page are displayed. The Report Header in the object only at the beginning of a report, the Detail in the object will show the results set in reporting, Report Footer in the object only in the end of the reporting shows once. We placed the field need to display in the Details area, through the Structure tab to see the composition relationship between reporting data items and pages.

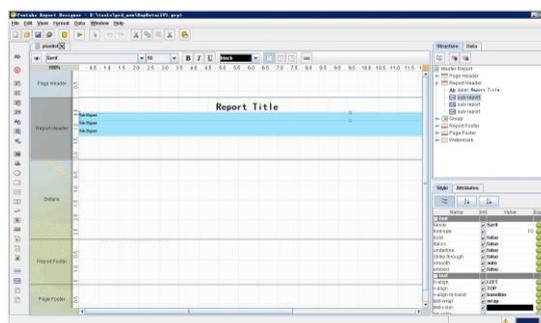


Figure 4 Pentaho Designer Report design

Report Designer reporting can set the input parameters, through the parameters on reporting data source filter can pass interactive query conditions. Specific steps are as follows:

- Select parameter in the Data tab, click add or modify Data Sources from the right-click menu.
- Add a new query, named Query1.
- Named Start Date for parameter, designated query for Query1.
- Start Date parameter was added in the main query reporting query: Start Date parameter format is \$ {Start Date}. When the reporting displays train\_date = Start Date parameter, the query statement chosen field.

Running this report, a page with tip form will display.

**4.4 Custom reporting controls.** One of the biggest advantages of open-source business intelligence is custom the reporting control, to expanding development tools. In the development of the train compile information management, it is necessary to achieve sub-seats respectively, the designated area, the train grades and direction of the train set and the actual ability query to get the statistics of passenger train-couple table and capacity, furthermore, compile information, cross-road information, and basic information about the train running mileage, the number of vehicle group, compile layout, cross-road graphics, train running rules and so on. In Pentaho, we use JfreeChart to implement a custom report controls[6], and reused in other reporting through drag and drop controls.

**4.4.1 meta-elements.xml.** Definition of element definition in meta-elements.xml:

```
<meta-data>
<include-globals/>
<element name="station" hidden="false" bundle-name="metadata" implementation="Station Chart
Type">
    <attribute-group-ref ref="common-attributes" />
    <attribute-group-ref ref="interactivity" />
    <attribute name="start-station" mandatory="true" hidden="false" value-type="java. lang.
String"/>
    <attribute name="compile" mandatory="true" hidden="false" value-type="java. lang.
String"/>
    <!-- other attributes/>
</element>
</meta-data>
```

**4.4.2 StationChartType.java.** Implement the following control operations in the StationChartType.java file:

```
public class Station Chart Type implements Element Type
{
    //Control properties
    private String compile;
    private float couple;
    // Attributes defined in the meta-elements.xml file
    private transient Element Meta Data element Type;
    //Default constructor
    public Station Chart Type () {}
    //load the default metadata about the star element type
    public Element Meta Data get Meta Data () { /* Get metadata information */}
    public Object get Value (final Expression Runtime runtime, final Element element) { /* Get the
reporting interface custom value */}
    XY Dataset create Dataset () throws Parse Exception { /* Get data source */}
    JFree Chart create Chart (XY Dataset param XY Dataset) throws Parse Exception {JFree Chart
local JFree Chart = Chart Factory. Create Time Series Chart (null, "TIME", null, param XY Dataset,
false, false, false);
...../*Create chart */ return local JFree Chart;}
    // Setter function control properties
    void setCompile (String p1) {this.compile = p1;}
    void setCouple (float p2) {this.couple = p2;}}
```

The entire package of files compiled into a jar package put into PRD lib directory, restart the PRD will see a red bar on the left controls for the locomotive StationChart custom chart, and user can drag it to the reporting.

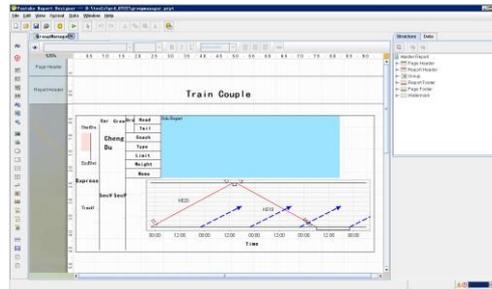


Figure 5 Customed Control in Pentaho Report Designer

## 5 Conclusion

This paper presents a method of constructing Pentaho business intelligence system in the Railway Passenger Marketing System, proposed ETL of Kettle design, based on reporting cube design, OLAP analysis, reporting and personalized report controls and other solutions according to user need to customize, solved the code closure problem of commercial software, while saving the cost of purchasing the software, so we conclude above open-source business intelligence projects have unique advantages. Though Pentaho platform integrates a set of excellent development, deployment and operation environments and tools of business intelligence systems, enterprise-class business intelligence implementation provides a good platform, in the development, we also found that of open-source business intelligence platform also has some disadvantages: including the development of the heavy workload, lack of software documentation and other issues more prominent, large-scale production and the security of the system did not do a very good concurrent performance testing and optimization, which is the most important issue in future research and production need to be addressed.

## References

- [1] J.X.Wang,C.H.Liu,J.F.Zhang,Application of Business Intelligence technology to Railway Passenger Transport Marketing Aided Decision System. *Railway computer application*.12,23 (2009)
- [2] Pentaho Corporation . Pentaho open-source business intelligence platform technical white paper. <http://www.pentaho.com>(2015)
- [3] Pentaho Documentation Team. Evaluate and Learn Pentaho Data Integration. <https://help.pentaho.com/Documentation/5.3/0D0/1A0/010/000>, (2015)
- [4] Rong-xin Chen,Y.G.Fu,W.B.Chen,Business intelligence system based on Pentaho.Computer Engineering and Design. 9.2408. (2008).
- [5] jfree.org . JFreeChart API Documentation. <http://www.jfree.org/jfreechart/api/javadoc/index.html>.(2015)
- [6] Pentaho Documentation Team. Business Analytics Report Designer. <http://www.pentaho.com/training-course/business-analytics-report-designer>.(2015)