# R role in Business Intelligence Software Architecture

**Ettore Colombo**[1,*], **Gloria Ronzoni**[1], **Matteo Fontana**[1]

1. CRISP (Interuniversity Research Center on Public Services), University of Milan Bicocca
*Contact author: ettore.colombo@crisp-org.it

**Keywords:** Business Intelligence, Data warehousing, Pentaho, **Rserve**

During the last years, design and development of Business Intelligence Systems and Data Warehouse have become more and more challenging. The reason why mainly stands in the continuos growing of both the amount of data to elaborate and the need of more effective models for data analysis and presentation.

The current situation has brougth to the definition of new software architectures aiming to reduce the computational burden and which, at the same time, could preserve qualities of reliability, accessibility, scalability, integrability, soundness and completeness of the final application.

The work hereby presented aims to show the solution adopted at CRISP (Interuniversity Reseach Center on Public Services) and how R has been adopted. R has been integrated in the suite of tools exploited to build a Decision Support System aiming at "knowledge workers"in the public sector. For example, end-users of this system are people like decision makers and data analysts working on service and policy design in Italian local government. The technological solutions adopted at CRISP are based on the integration of different tools coming from the Open Source community. This suite is composed by software tools that cover all the layers that are required in building a data warehouse system. *Data Transformation and Preparation* are perfomed using Talend Open Studio, an open source ETL (Extraction, Transformation and Loading) and data integration platform, *Data Storing* is obtained exploiting MySQL, the well-known open source DBMS (Data Base Management System), while Pentaho BI (the world leader open source Business Intelligence platform) is deputated to *OLAP* (On-Line Analysis Processing) and *Data Presentation* (e.g. reporting and dashboarding).

The role of R in this architecture is twofold. On the one hand, R can play a central role in data elaboration according to innovative or well-known analysis model when interfaced to Talend Open Studio during ETL processes. For instance, running R directly via command-line within Talend Open Studio, it is possible to apply specific models to analyze data stored in MySQL databases and directly modify them (e.g. to determine clusters for employee careers classification).

On the other hand, R has been integrated in the Data Presentation layer by means of **Rserve**, a package that makes R accessible via TCP/IP. A Pentaho BI component that extends the existing platform, called RComponent, has been directly developed at CRISP in order to manage communication issue between R and Pentaho BI.

In particular, the obtained integrated suite has been used in *Labor*, a project aiming to analyze the labour market in different Italian provinces of Regione Lombardia in order to classify employee careers and provide the system with advanced data presentation solutions based on complex statistical models (e.g. Markov chains for predictive models).

## References

Golfarelli M. , Rizzi S. (2002) Data Warehouse. McGraw Hill.

Rserve project Home Page
`http://rosuda.org/Rserve/`.

Pentaho Home Page
`http://www.pentaho.com/`.