A Perspective on the Benefits of Data Virtualization Technology

Ana-Ramona BOLOGA, Razvan BOLOGA
Academy of Economic Studies, Bucharest, Romania
ramona.bologa@ie.ase.ro, razvanbologa@ase.ro

Providing a unified enterprise-wide data platform that feeds into consumer applications and meets the integration, analysis and reporting users' requirements is a wish that often involves significant time and resource consumption. As an alternative to developing a data warehouse for physical integration of enterprise data, the article presents data virtualization technology. There are presented its main strengths, weaknesses and ways of combining it with classical data integration technologies. Current trends in data virtualization market reveal the great potential of growth of these solutions, which appear to have found a stable place in the portfolio of data integration tools in companies.

Keywords: *Data virtualization, Data service, Information-as-a-service, Integration middleware*

1 Introduction

Present paper focuses on the role of data integration process in the business intelligence projects and on how data virtualization technology can enhance this process, extending the conclusions presented in [1]. Business intelligence includes software applications, technologies and analytical methodologies that perform analysis on data coming from all the significant data sources of a company. Providing a consistent, single version of the truth coming from multiple heterogonous sources of data is one of the biggest challenges in business intelligence project.

The issue of data integration effort is well known in the field. Almost every company manager that was involved in a Business Intelligence projects during the last years accepts there were data related problems that generated a negative effect on business and extra-costs to reconciliate data.

"Anecdotal evidence reveals that close to 80 percent of any BI effort lies in data integration... And 80 percent of that effort (or more than 60 percent of the overall BI effort) is about finding, identifying, and profiling source data that will ultimately feed the BI application"[5].

Experience reveals that "more than technical reasons, organizational and infrastructure dysfunction endanger the success of the project" [7].

The already classical solution to data integration problem in Business Intelligence consists in developing an enterprise data warehouse that should store detailed data coming from the relevant data sources in the enterprise. This will ensure a single view of business information and will be the consolidated data source used further for dynamic queries and advanced analysis of information.

Though, building an enterprise data warehouse is a very expensive initiative and takes a long time to implement. Data warehouse based approach has also important constraints that drawbacks its appliance to highly decentralized and agile environments. What are the alternative approaches when you don't have the appropriate budget or when you need a fast solution?

During the last years the data virtualization concept gained more and more adepts and data virtualization platforms were developed and spread over. What is the coverage of this concept, when and where it can be used is the subject of the following paragraphs.

2 Data virtualization

Data virtualization is "the process of abstraction of data contained within a variety of information sources such as relational databases, data sources exposed through web services, XML repositories and others so that they may be accessed without regard to their physical storage or heterogeneous structure" [11]. Data virtualization is a new term and used especially by software vendors, but it hides the old idea of data federation as an important data integration technology.

It can be used in a variety of situations that require unified access to data in multiple systems via high-performance distributed queries. This includes data warehousing, reporting, dashboards, mashups, portals, master data management, SOA architectures, post- acquisition systems integration, and cloud computing [3]. Of course the tools used today for data federation are much more powerful and the technologies used are different but the idea was used much before.

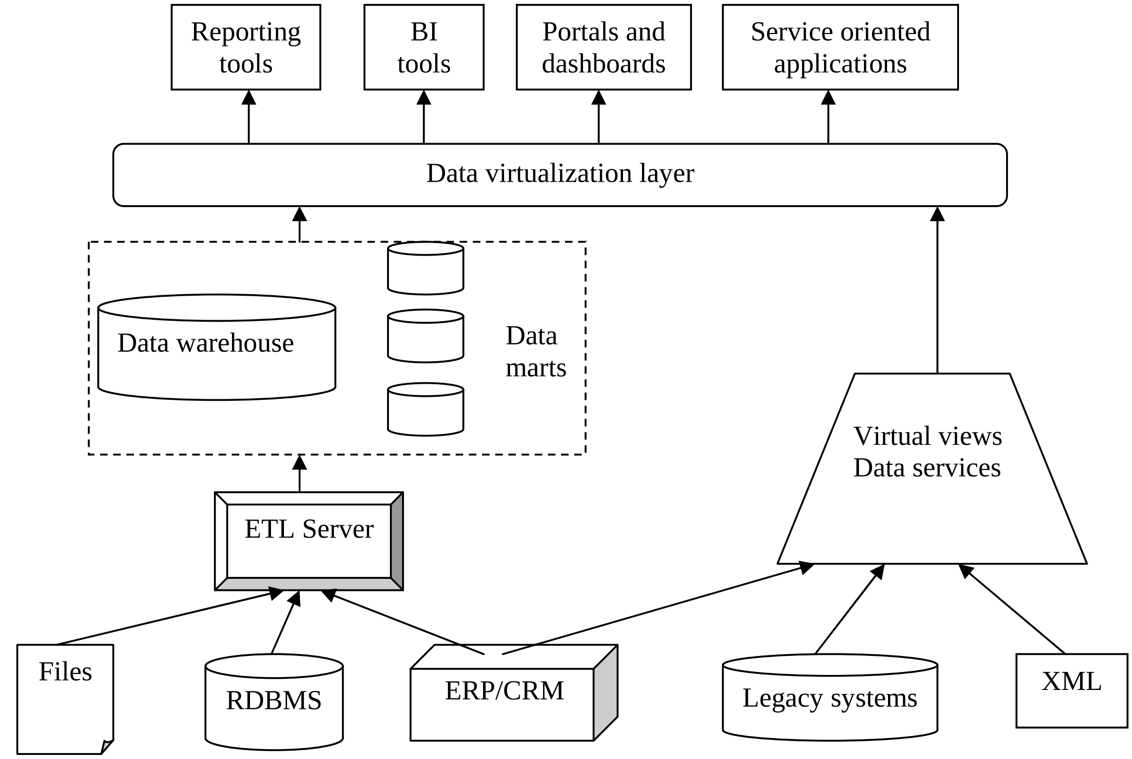

Fig. 1. Data virtualization

In ‘80s, database vendors developed database gateways for working with ad-hoc queries on multiple heterogonous databases.

In ‘90s, software vendors tried to create virtual data warehouses applying data federation. At the time, the integration of data coming from legacy systems was almost impossible without using a new data store.

In 2000s, Enterprise Information Integration (EII) proposes an architecture that allows collection of data from a various data sources for providing some reporting and analysis applications with the required information. This architecture focuses on viewing the right information in the right form by moving only small pieces of data. EII was not scalable, it was designed to offer point-to-point integration, so once two applications were integrated and a new application is added in the system, the integration effort must be repeated.

In the recent years data federation has also been called data services or information as a service. Service oriented architectures use data federation as a data service that abstracts back-end data sources behind a single query interface. Data services offer a flexible data integration platform based on the new generation of standards or services that allows access to any data type, located on any platform, using a great variety of interfaces and data access standards. But data services can offer more than that: support for a single version of the truth, real time business intelligence, searching throughout all company’s data, high level security in accessing sensitive data.

Moreover, you can bind the virtualized data to additional function behavior, and thus create very powerful reusable data services that again can be mixed and matched to create solutions [6].

Advanced data virtualization brings an important progress in power and capabilities. It abstracts data contained in a variety of data sources (databases, applications, data warehouses) and stores the metadata in a persistent metadata directory in order to offer a single data access point.

Data virtualization involves capabilities like:

- Abstracting of information related to the technical aspects of stored data, like location, storage structure, access language, application programming interfaces, storage technology and so on. Data are presented to consumer applications in a consistent way, regardless of native structure or and syntax that may be in use in the underlying data sources.
- Virtualization of data sources (databases, Web content and various application environments) connection process, making them logically accessible from a single point, as if they were in one place so that you can query that data or report against it.
- Transformations for improving data quality and integration for collecting data from multiple distributed sources.
- Federation of data sets coming from various, heterogonous source systems (operational data or historical data or both).
- Flexibility of data delivery as data services are executed only when users request them.
- Presentation of the required data in a consistent format to the front-end applications either through relational views or by means of Web services.

Also, data virtualization capabilities delivery comes together with capabilities for data security, data quality and data management requirements, for queries optimization, caching, and so on.

Data virtualization process involves two phases:

1. Identification of data sources and of the attributes that are required and available for the final application. If there are more than one data source that can provide the required data, the analyst will decide which of the sources is more trusted and will be used in building the response. Data virtualization tool will be used for designing a *data model* that defines the entities involved and creates physical mappings to the real data sources.

Advanced data virtualization tools will create a business object model as they use object oriented modeling for internal representation of every stored data structure (even the relational data) which enhances many applications through the power of objects.

2. Application of the data model in the second step for getting data from various data sources in real time when there is a client application query. The language used for query description can be any standard language used by the data virtualization tool.

3 Federation vs. Integration

The traditional way of integrating data in business intelligence project consisted in developing a data warehouse or a collection of data marts. That involved designing new data storage and using some ETL tools (Extract, Transform and Load) to get the required data from the source systems, clean them, transform them to satisfy the constraints of the destination system and the analysis requirements and, finally, load them into the destination system. Usually, developing an enterprise data warehouse takes at least several months and involves important financial and human resources, but also a powerful hardware infrastructure to support the storage and the processing in seconds of terabytes of information. Data integration has also some specific advantages: data quality and cleaning, complex transformations.

Data virtualization software uses only metadata extracted from data sources and physical data movement is not necessary. This approach is very useful when data stores are hosted externally, by a specialized provider. It allows real-time queries, rapid data aggregation and organization, complex analysis, all this without the need of logical synchronizations or data copying.

It is better to use to use virtualization tools to build a virtual data warehouse or to use ETL tools to build a physical data warehouse?

Data virtualization is recommended especially for companies that need a rapid solution but does not have the money to spend for consultants and infrastructure needed by a data warehouse implementation project.

Using data virtualization the access to data is simplified and standardized and data are retrieved real-time from their original sources. The original data sources are protected as they are accessed only trough integrated views of data.

But data virtualization is not always a good choice. It is not recommendable for applications that involve large amounts of data or complex data transformation and cleaning, as those could slow down the functioning of source systems. It is not recommended if there is not a single trusted source of data. Using unproven and uncorrected data can generate analysis errors that influence decision making process and can generate important losses for the company.

But, data virtualization can also complement the traditional data warehouse integration. Here are some examples of ways of combining the two technologies whose practical application has proved to be very valuable [12]:

a. **Reducing the risk of reporting activity during data warehouse migration or replacement** by inserting a virtual level of reporting between data warehouse and reporting systems. In this case, data virtualization makes it possible to continue using data for reporting during the migration process. Data virtualization can help reducing costs and risks by rewriting report queries for data virtualization software instead of old data. When the new data source is ready, the semantic layer of data virtualization tool is updated to point the new data warehouse.

b. **Data preprocessing for ETL tools** as are not always the best approach for loading data into warehouses. They may lack interfaces to easily access data sources (e.g. SAP or Web services). Data virtualization can bring more flexibility by developing data views and data services as inputs to the ETL batch processes and using them as any other data source. These abstractions offer the great advantage that ETL developers do not need to understand the structure of data sources and can be reused any time it is necessary. Virtual views and data services reuse lead to important cost and time savings.

c. **Virtual data mart creation** by data virtualization which significantly reduce the need for physical data marts. Usually, physical data marts are built around data warehouse to meet particular needs of different departments or specific functional issues. Data virtualization abstracts data warehouse data in order to meet consumer tools and users integration requirements. A combination of physical data warehouse and virtual data marts could be applied to eliminate or replace physical marts with virtual ones, such as stopping rogue data mart proliferation by providing an easier, more cost-effective virtual option.

d. **Extending the existing data warehouse** by data federation with additional data sourcing, also extending data warehouse schema. Complementary views are created in order to add current data to the historical data warehouse data, detailed data to the aggregated data warehouse data, external data to the internal data warehouse data.

e. **Extending company master data**, as data virtualization combines master data regarding company’s clients, products, providers, employees and so on with detailed transactional data. This combination brings additional information to allow a more comprehensive view of company activity.

f. **Multiple physical data warehouse federation** as data virtualization realizes logical consolidation of data warehouses by creating federated views to abstract and rationalize schema designs differences.

g. **Virtual integration of data warehouse**

in Enterprise Information Architectures which represents the company's unified information architecture. Data virtualization middleware forms a level of data virtualization hosting a logical scheme that covers more consolidated and virtual sources in a consistent and complete way.

h. **Rapid data warehouse prototyping** as they data virtualization middleware serves as prototype development environment for a new physical data warehouse. Building a virtual data warehouse leads to time savings compared to the duration involved in developing a real data warehouse. The feedback is quick and adjustments can be made in several iterations to complete a data warehouse schema. The resulted data warehouse can be used as a complete virtual test environment.

So, data visualization represents an alternative to physical data integration for some specific situation, but can always come and complement the traditional integration techniques.

Could these solutions be somehow combined in a single one, so we get both sets of advantages? The answer is positive, if there was a way of using the semantic model from data virtualization for applying ETL quality transformations on real time data. This would mean a single integration toolset integrating both virtualization and data integration capabilities. This is the solution that integration vendors are reaching after.

4 Data virtualization tools

4.1 Market analysis

The data virtualization market currently stands at around $3.3 billion and it will grow to $6.7 billion by 2012, according to Forrester Research [8].

The Forrester study of data virtualization tools market indicates as top vendors Composite Software, Denodo Technologies, IBM, Informatica, Microsoft, and Red Hat. These top vendors can be grouped in two categories: large software vendors (IBM, Informatica, Microsoft, and Red Hat) which offer broad coverage to support most use cases and data sources and smaller, specialized vendors (Composite Software, Denodo Technologies, and Radiant Logic) which provide greater automation and speed the effort involved in integrating disparate data sources.

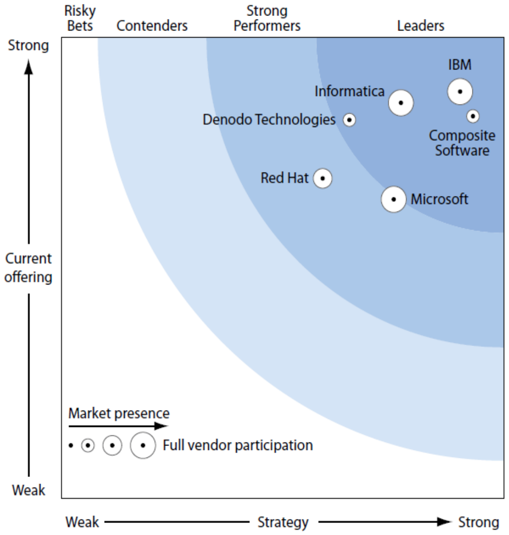

Fig. 2. Data virtualization tool vendors [8]

Other BI vendors support data federation natively, including Oracle (OBIEE) and MicroStrategy. Most popular integration tools providing data federation features are:

- SAP BusinessObjects Data Federator;
- Sybase Data Federation;
- IBM InfoSphere Federation Server;
- Oracle Data Service Integrator;
- SAS Enterprise Data integration Server.

Data virtualization options offered by these tools are much diversified, including:

- federated views;
- data services;
- data mashups;
- caches;
- virtual data marts;
- virtual operational data stores.

They support a range of interface standards, including REST, SOAP over HTTP, JMS, POX over HTTP, JSON over ODBC, JDBC, ADO.NET and use many optimization techniques to improve their performances, including rule-based and cost-based query-optimization strategies and techniques: parallel processing, SQL pushdown, distributed joins, caching or advanced query optimization.

An analysis of the companies implementing or intending to implement data integration solutions reveals some dominant market trends [4][5]:

a. Maximum exploitation of the already existing technology, as many companies do not wish to create a special platform for data integration and prefer to use the existing components and processes for starting such an initiative.
b. Focus on read-only use cases, as most developers use this approach delivering virtualization solutions for real time business intelligence, a single version of the truth providing, federated views and so on. Still, the offer of read-write virtualization solutions has increased in recent years.
c. An increasing interest in using cloud computing, as most of data services providers already offer REST support, allowing “on-premise” integration. For the moment, there were used basic cloud-based services for solving simple integration tasks for organizations with restricted resources. But big organizations are also interested in cloud-based infrastructures as a way of offering non-production environments for the integration solutions they use.
d. Data integration tools market and data quality tools market convergence is more accelerated. Both types of tools are necessary for initiatives in business intelligence, master data management or application modernization. Most companies that have purchased data integration tools have also wanted data quality features.

The following section will briefly present a data virtualization solution in order to capture the complexity and main capabilities of such solutions and to illustrate the importance of its integration with other tools for working with integrated data at enterprise level.

4.2 Composite Software solution

Composite Software solution seems to be one of the most complete and competitive solutions for data virtualization available on the market on this moment. This fact has also been reflected in the analysis realized by the Forrester Group described above. At least, Composite is the leader pure play vendor in the marketplace.

Composite Software was founded in 2002 and was for many years a niche player with its specialized product for data virtualization. The Composite Data Virtualization Platform is data virtualization middleware that ensures a unified, logical virtualized layer for disparate data sources in the company.

Composite platform is based on views, which are relatively simple to define. After defining views it can automatically generate SQL code or WSDL code. It offers pre-packaged interfaces for the leading application environments in order to simplify the integration effort. It also provides some very useful out-of-the-box methods, such as SQL to MDX translators and methods for simplified access to Web content.

It offers support for SQL and XML environments but also for legacy environments, flat files, Web content and Excel.

Its architecture, presented in Fig. 3, has two main components, both of which are pure Java applications, delivered as Linux based appliance, with browser based clients:

- **Composite Information Server** includes a query engine with Query optimization technology, caching, data access, data quality, security, data governance and a metadata repository;
- **Composite Discovery** is an enterprise data entity and relationship discovery tool used for preparing data models and performing intelligent queries that is now tightly integrated with Composite Information Server.

In addition, other important components are:

- **Composite Monitor** includes an integrated development environment for source data introspection, relationship discovery, data modeling, view and data service development, and revision control. Introspection facility is very useful to detect the made customizations that can be used for appropriate adjustments of standard interfaces.

- **Composite Active Clusters** allows multiple instances of Composite Server deployment for scalability and high availability. For shared metadata caching can be used if Active Clusters is implemented together with an environment as Oracle RAC (Real Application Clusters).
- **Composite Application Data Services** contains packages of services to connect to the main environments applications (for Oracle, Oracle Essbase, Salesforce.com, SAP, SAP BW, and Siebel). If working with MDX data sources (Oracle Essbase and SAP BW), the service automatically convert SQL queries based on the MDX. It offers a graphical OLAP Viewer which displays multidimensional databases as relational structures, making them easier to understand to SQL developers.

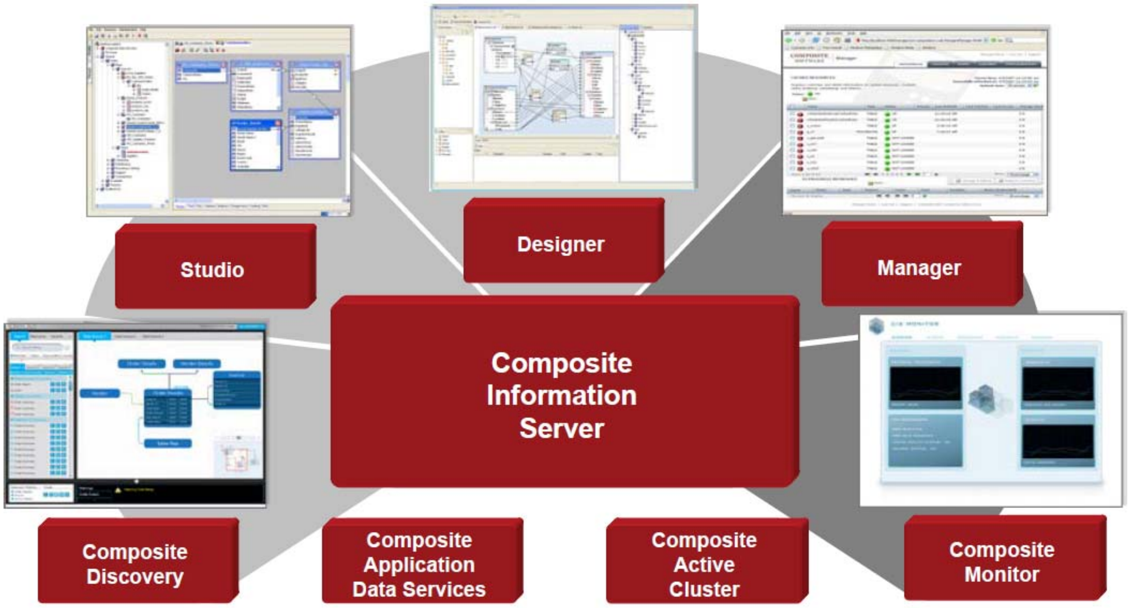

Fig. 3. Composite Data Virtualization Platform [9]

Composite Software focuses on obtaining optimal performance in data integration. The query processing engine and the cache are the two pillars that enable performance optimizations. The query processing engine works using some query plans that aim to reduce network traffic for distributed queries. For this, most of the processing work should be done in the source databases. There is an optimizer that uses rules and information on cost and network load to determine the best place for carrying out operations.

The cache usage makes query processing faster and it is completely under user control. The user decides its dimension, the refreshment mode, and when data is purged and can define customized indexes.

Data sources supported include:

- Flat files;
- XML data sources (XQuery support);
- Any JDBC compliant sources;
- MS Excel through ODBC;
- DB2 on zOS;
- Composite Information Server data;
- Legacy mainframe sources by reselling Data Sirect Shadow RTE software.

Specific integration capabilities have been developed in collaboration with Netezza, taking advantage of the host environment for performance purposes. In fact, this partnership approach seems to have an important contribution to the company's upward trend in recent years. Composite Software has signed an OEM agreement with Cognos in 2004 and Composite Information server is embedded in Cognos. A number of other business intelligence software vendors resell Composite Application Data Services. There was also collaboration with Kapow Technology to develop data and complement the native Composite Application Data Services for Web content.

It seems that collaborations and partnerships, openness to others expertise is the best solution for developing a complete and competitive data virtualization solution, as it have had a positive impact on both sales and the development of functionality and coverage of Composite Software solution.

5 Conclusion

Efficient, competitive decisions in real-time, with minimum cost and investment is the dream of every modern manager. The economic conditions and opportunities are very dynamic, so an efficient business intelligent solution has to meet the ever-changing information needs of today's enterprises. The advantages offered by data virtualization are very appealing: reduced costs, reduced time to implement, greater agility, and reuse.

Data federation is a very important and promising tool which can be used independently or complementary to physical data integration. But there are specific use-cases for which the use of data virtualization is recommendable.

A competitive advantage can be obtained only by adding business-oriented personalization so that the information provided can fulfill the particular needs of end-users and empower dynamic analysis and decisions [2].

For the case of business intelligence, the decision of adopting a virtualized solution should be carefully analyzed. BI projects usually involve complex multidimensional analysis and a great importance is given to data cleaning and consolidation. The number and dimension of data sources, the quality of the raw data and the analysis requirements should also be considered, as they have an important impact on data integration and may entirely compromise the advantages of virtualization technology.

For the moment, the optimal solution seems to be the adoption of a portfolio of data integration tools that should respond to various integration needs and should support the data access requirements.

In many cases, the best integration solution is a combination of virtual and physical approaches, keeping the physical data warehouse in order to benefit from its features and applying virtualization for cutting costs and getting quicker results for data source access, for data mart elimination, for prototyping new data marts or data warehouses, for federating multiple physical consolidated sources and so on.

References

[1] A. R. Bologa, R. Bologa, “Data Virtualization – Solution for Data Integration in Business Intelligence Projects,” *Proceedings of The Tenth International Conference on Informatics in Economy IE 2011*, May 5-7, 2011, ISSN 2247-1480, ISSN-L 2247-1480;

[2] A. R. Bologa, R. Bologa, A. Bara, “Technology vs Business Needs in Business Intelligence Projects,” *Proceedings of the International Conference on e-business (ICE-B)*, 2008, ISBN 978-989-8111-58-6;

[3] W. Eckerson, *The Evolution of Data Federation*, http://www.massivedatanews.com/ content/evolution-data-federation, June 2010;

[4] T. Friedman, M. A. Beyer, E. Thoo, “Magic Quadrant for Data Integration Tools,” *Gartner RAS Core Research Note G00207435*,19 November 2010;

[5] H. Kisker, J.P. Garbani, B. Evelson, C. Green, M. Lisserman, “The State of Business Intelligence Software and Emerging Trends, 2010,” *Forrester Research report,* May 10, 2010, http://www.forrester.com/rb/Research/state_of_business_intelligence_software_and_emerging/q/id/56749/t/2;

[6] D Lithicum, “Rethinking Data Virtualization,” *Informatica Perspectives*,

http://blogs.informatica.com/perspectives/ index.php/2010/04/21/rethinking-data-virtualization, April 2010;

[7] A. R. Lupu, R. Bologa, I. Lungu, A. Bara, "The Impact Of Business Organization On Business Intelligent Projects," *Proceedings of the 7th WSEAS Int. Conf. on Simulation Modeling and Optimization*, pp. 415-419, Beijing, China, 2007, ISBN, 978-960-6766-05-3;

[8] N Yuhanna, MGilpin, *The Forrester Wave™: Information-As-A-Service*, February 2010, http://www.informatica.com/ downloads/7015_forrester_wave_iaas_ 2010.pdf

[9] *Composite Software Products,* http://www.compositesw.com/products/

[10] *Data Federation Solutions: Accelerate BI Projects spanning Disparate Data Sources*, file:///H:/conference%20ie%202011/Composite%20 Software%20_%20Data%20 Federation.htm

[11] *Data Virtualization*, http://en.wikipedia.org/wiki/Data_ virtualization;

[12] *Twelve Key Reasons to Use Composite Data Virtualization*, Composite Software, January 2010, http://purl. manticoretechnology.com/ImgHost/582/12917/2011/ Document_ Downloads/12Reasonsfor Composite DV.pdf

Ana-Ramona BOLOGA (born in 1976) is lecturer at the Academy of Economic Science from Bucharest, Economic Informatics Department. Her PhD paper was entitled "Software Agents Technology in Business Environment". Her fields of interest are: integrated information systems, information system analysis and design methodologies, and software agents.

Razvan BOLOGA (born 1976) is lecturer at the Academy of Economic Studies from Bucharest Romania. He is part of the Economic Informatics Department and his fields of interest include information systems, knowledge management and software ecosystems. Mr. Bologa has attended over 15 conferences presenting the results of his research