

## **OS BI: Open Research Issues**

Authors:

Matteo Golfarelli<sup>1</sup>, Stefano Rizzi<sup>1</sup>, Yannis Velegrakis<sup>2</sup>

<sup>1</sup>*DEIS - University of Bologna - Italy*

<sup>2</sup>*DISI, University of Trento - Italy*

This work is licensed under the Creative Commons Attribution-Share Alike 3.0 License.

To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/3.0/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 1/18

**Revision History**

Version	Lead Author	Summary of Changes	Date
1.0	DEIS, Univ. BO, DISI, Univ. TN	First version release	Dec. 22 <sup>nd</sup> , 2008

**Table of Contents**

Abstract ..... 3

About the BI Initiative ..... 4

1. Introduction ..... 5

2. Strengthening the first BI generation ..... 8

    2.1 Design ..... 8

    2.2 Evolution and versioning ..... 9

    2.3 User preferences ..... 9

    2.4 What-if analysis ..... 10

3. Towards the second BI generation ..... 11

    3.1 Pervasive BI ..... 11

    3.2 Distributed BI ..... 12

    3.3 Open Source Mapping Systems and Benchmarks ..... 13

Conclusions ..... 16

References ..... 16

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 2/18

## Abstract

Business Intelligence can be defined as the process of turning data into information and then into knowledge. Business Intelligence systems are centered on data warehouses: large repositories of historical data, organized according to the multidimensional model. In spite of the large amount of research carried out in the last 20 years, the modern end-to-end business intelligence platforms seem to suffer from the increased/changed user needs. This is a consequence of the growth of the business intelligence culture within enterprises on the one hand, and a side effect of the mutated business panorama on the other. While in the past the business intelligence market was strictly dominated by closed source and commercial tools, the last few years were characterized by the birth of open source solutions. Commercial platforms are commonly considered superior to open source ones. Nevertheless, open source platforms will evolve much faster than commercial platforms since they are not constrained by compatibility problems and rigid (or even obsolete) architectures.

In this paper we will propose a list of relevant research issues related to data warehouse and business intelligence and we will discuss how and why they can be used to create the second generation of business intelligence tools. In particular, we will distinguish between mid and long term goals and we will envision some possible applications.

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 3/18

## About the BI Initiative

OW2 Initiatives are joint efforts of OW2 Members and not-OW2 Members aiming at facilitating the use of OW2 technologies by mainstream systems integrators, end-users, academia and software vendors. Within an Initiative, participants work together to develop both technical integration between projects and business synergies in order to address specific market needs.

The BI Initiative is a join effort set up to:

- improve the coordination effort in the OS BI context
- increase the use of OS BI solutions at enterprise level
- strengthen connections between integrators, vendors, users and the research communities
- attract more attention from the research activities to foster innovative BI solutions and practices.

Main activities are:

- integration of a full stack of OS BI solutions and tools
- promotion of a service network to support the entire stack
- kick-starting a community contributing to strengthen the current OS BI platforms
- creation of a research network on BI topics, fostering a closer cooperation between researchers and OS BI vendors, to promote the development of a new generation of OS BI platforms.

Participants of the BI Initiative are:

- Altic, France ([www.altic.org](http://www.altic.org))
- Artemis Information Management, Luxembourg ([www.artemis.lu](http://www.artemis.lu))
- ClaraVista, France ([www.claravista.fr](http://www.claravista.fr))
- Engineering Ingegneria Informatica, Italy ([www.eng.it](http://www.eng.it))
- Ingres, USA ([www.ingres.com](http://www.ingres.com))
- Talend, France ([www.talend.com](http://www.talend.com));
- DEIS-University of Bologna-Italy ([www.eng.unibo.it/PortaleEn/default.htm](http://www.eng.unibo.it/PortaleEn/default.htm))
- DTI-University of Milan (<http://ra.crema.unimi.it>), Italy
- DISI-University of Trento ([www.dit.unitn.it](http://www.dit.unitn.it)), Italy
- OW2 individual members.

More information about OW2 BI Initiative: [www.ow2.org/view/BusinessIntelligence/](http://www.ow2.org/view/BusinessIntelligence/)

More information about OW2 Consortium: [www.ow2.org](http://www.ow2.org)

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 4/18

## 1. Introduction

Business Intelligence (BI) can be defined as *the process of turning data into information and then into knowledge*. BI was born within the industrial world in the early 90's, to satisfy the managers' request for efficiently and effectively analyzing the enterprise data in order to better understand the situation of their business and for improving the decision process. In the mid-90's BI became an object of interest for the academic world, and ten years of research managed to transform a bundle of naive techniques into a well-founded approach to information extraction and processing.

In particular, a huge work has been done in the specific area of *data warehouses* (DWs). DWs are large repositories of historical data, organized according to the multidimensional model. They are directly accessed by final users (i.e. the managers) through user-friendly interfaces that enable them to carry out very detailed analyses. The main results obtained on topics such as OLAP visualization [11], multidimensional modeling [4], design methodologies [5] and optimization techniques [12] converged to define the modern architecture of data warehousing systems and were absorbed by vendors to form a wide set of on-the-shelf software solutions. Today, data warehousing can be considered a mature field from several points of view: users understood the potential of multidimensional analysis and are fully exploiting OLAP capabilities; complete suites of tools are available and cover the whole data warehousing process from ETL extractors, to friendly interfaces, through specialized DBMSs; researchers explored most issues related to the conceptual, logical and physical levels of DW architectures.

In spite of this, the modern end-to-end BI platforms seem to suffer from the increased/changed user needs. This is a consequence of the growth of the BI culture within enterprises on the one hand, and a side effect of the mutated business panorama on the other. The managers, aware of the BI capabilities and pushed by a more and more complex and frantic market, are asking for new solutions made feasible by the technology improvements. Beside OLAP, data mining and what-if analysis solutions must be made easily available to the users in the next years possibly in a unified and transparent fashion. Some proposals in this direction already coined the term OLAM (OnLine Analytical Mining model). Basically, this proposal consists in meaningfully combining the powerful of OLAP with the effectiveness of Data Mining tools and algorithms capable of

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 5/18

discovering interesting knowledge from large amounts of data (e.g., the data cell set of a given OLAP data cube) by means of clustering, classification, association rule discovery, frequent item set mining, and so forth.

While in the past the BI market was strictly dominated by closed source and commercial tools (see for example [13] for different vendors' market shares of OLAP servers), the last few years were characterized by the birth of open source (OS) solutions. In the beginning OS BI tools covered isolated portion of the DW process with a limited set of functionalities that made them appear as toys if compared to large commercial BI platforms. Consider for example the initial releases for Octopus as to ETL, Mondrian as to OLAP servers, and JPivot as to OLAP clients (see [16] for a complete listing). While single tools still keep evolving with an increasing number of features and a higher level of reliability, the turning point in OS BI has been the birth of OS BI platforms. An OS BI platform provides a full spectrum of BI capabilities within a unified system that reduces the overhead for the development and management of each application, and lets the user feel like she was using a single BI solutions.

Commercial platforms are commonly considered superior to OS platforms. Nevertheless, OS BI platforms will evolve much faster than commercial platforms since they are not constrained by compatibility problems and rigid (or even obsolete) architectures. Furthermore, OS solutions can exploit the contributions of the OS development community, that relies on hundreds of programmers and designers as well as on the direct involvement of researchers that are willing to prototype on free platforms.

In order to really do better than commercial solutions, we argue, OS solutions should not only replicate commercial functionalities with lower costs for the final users, but should also propose innovative functionalities according to the most sophisticated requirements of business users. Coupling twenty years of experience in building BI software with the more recent results on BI research can really make the difference. Thus, in this paper we list and discuss some research topics that are currently being investigated and that, we believe, could suggest new functionalities to be added to OS BI platforms. We envision two main directions: on the one hand, BI platforms need to be consolidated since they still lack in supporting some relevant requirements; on the other hand, part of the new user requirements impose to enlarge the traditional vision of BI since they cannot be fulfilled within the current frameworks. As a consequence, we should begin to work towards a

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 6/18

second generation of BI tools that, keeping into account the jumps ahead of both technology and research, will be capable of supporting new types of analysis.

Noticeably, a discussion on the second generation of BI started some years ago, during the Perspectives Workshop “Data Warehousing at the Crossroads” [10], in which the BI research community discussed the state of the art in the field and tried to anticipate the hot research issue for the short and medium term future. Some of the topics proposed at that time have already become object of interest for the market, while others still remain mostly unexplored. Two noticeable examples of topics that attracted the market attention are the following:

- *Spatial data warehousing.* Spatial and spatio-temporal data are inherently more complex than traditional (alphanumerical) business data. Spatial data refer to geometric objects like points, lines, regions, surfaces, and volumes in the two-dimensional and/or three-dimensional space. Spatio-temporal data add the temporal aspect. Besides, operations on geometric objects are much more complex than those on alphanumerical values. Spatial objects can change their position, shape, and extent over time and are then called moving objects. Geographical information systems (GIS) and data warehousing share the common interest to analyze data and to facilitate decision-making processes. But they use different technologies, and the capabilities of GIS for decision support are limited. Estimates state that 80% of all data have a spatial context or reference [8]. This makes the combination of both technologies an interesting and promising approach. Especially the support of ad hoc spatial and spatio-temporal aggregation operators is here of interest.
- *Data streams, real-time and active data warehouses.* Since DWs integrate data from different operational systems, they provide unique functionality for discovering information across these otherwise isolated data sources. This makes them ideal candidates for monitoring enterprise- or organization-level events that cannot be detected by a single operational system. Unfortunately there is a considerable delay until information from operational systems is reflected in the DW: updates are typically performed in large batches once a week or overnight. However, a bank or telephone company requires real-time data warehousing functionality to detect suspicious activity in customer accounts in a timely manner before it results in financial damages.

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 7/18

Similarly, a large retailer could better optimize the supply at geographically distributed stores and warehouses, if he had more timely information in the DW. Other examples include data warehousing solutions for data stream applications like stock trading, military logistics, traffic monitoring, and sensor networks in general. All these applications require real-time or nearly-real-time update propagation, hence also real-time ETL.

The above-mentioned topics require further investigations and a strong design and implementation effort for tools to fully exploit their potentialities, but they are on their way. In the following sections we will focus instead on those features that are currently not available at all in tools.

## 2. Strengthening the first BI generation

### 2.1 Design

The statistic reports related to DW project failures state that a major cause lies in the absence of a global view of the design process. Indeed, adopting a methodological framework for design is an essential requirement to ensure the success of complex projects. On the other hand, it is well-known among software engineers that devising a design methodology is almost useless, if no CASE tool to support it is provided.

Though some design methodologies for DWs have been proposed, none of them has been widely accepted so far, and all vendors propose their own proprietary methods. One of the reasons for this is that there is still no agreement about a standard conceptual model to be used at the core of the methodology. Conversely, a unified conceptual model for DWs would be a valuable support for both the research and industrial communities. It should be formally well-founded, but at the same time easily usable and understandable by designers. It should support integrated modeling of the DW architecture, deployment, sources, mappings, ETL, facts, workloads, etc. Finally, it should be expressive and flexible enough not only to enable representation of requirements coming from the classical enterprise domains, but also to support the peculiar issues and constraints arising in unusual and emerging domains and applications (such as those based on streaming data or geographical information).

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 8/18

## 2.2 Evolution and versioning

As several mature implementations of data warehousing systems are fully operational within medium to large contexts, the continuous evolution of the application domains is bringing to the forefront the dynamic aspects related to describing how the information stored in the DW changes over time. As concerns changes in data values, a number of approaches have been devised, and some commercial systems allow to track changes and to effectively query cubes based on different temporal scenarios [15]. Conversely, the problem of managing changes on the schema level (that may be required by changes either in the business domain or in the user requirements or in the sources) has only partially been explored, and no dedicated tools or restructuring methods are available to the designer yet.

The approaches to management of schema changes in DWs can be framed into two categories, namely evolution [1], [17] and versioning [2], [6]: while both categories support schema changes, only the latter keeps track of previous versions. If one is sure that previous schema information will never be useful again, schema evolution offers adequate functionality. Otherwise (e.g., to guarantee consistent re-execution of old reports), schema versioning offers the more powerful approach. Actually, in some versioning approaches, besides “real” versions determined by changes in the application domain, also “alternative” versions to be used for what-if analysis are considered [2].

Overall, we believe that versioning is better suited to support the complex analysis requirements of DW users as well as the DW characteristic of non-volatility. The main research challenges in this field are to provide effective versioning and data migration mechanisms, capable of supporting flexible queries that span multiple versions. Considering the complexity of the ETL procedures, another very relevant issue is to devise techniques for propagating changes occurred in the source schemata to the ETL process. The obvious benefit in achieving these goals will be to keep the DW in sync with the business requirements, thus avoiding its obsolescence.

## 2.3 User preferences

Personalizing e-services by allowing users to express preferences is becoming more and more common. When querying, expressing preferences is seen as a natural way to avoid empty results on the one hand, information flooding on the other. Besides, preferences allow for ranking query results so that the user may first see the data that better match her tastes. Though a lot of research

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 9/18

has been carried out during the last few years on database preferences (e.g., [3], [9]), the problem of developing a theory of preferences for multidimensional cubes has been mostly neglected so far [14]. On the other hand, cubes (implemented either on relational or on multidimensional platforms) are the core of data warehousing and business intelligence systems. Their users are decision makers who need to express complex queries through OLAP front-end tools, often returning huge volumes of data, sometimes returning little or no information. Thus, we argue that expressing preferences could be valuable in this domain.

What makes preference in OLAP systems different is the role played by aggregation, which enables decision makers to get valuable, summary information out of the huge quantity of transactional data available in the enterprise databases. OLAP queries do not only formulate selections and projections on the attributes and measures belonging to the cube, they also specify on what hierarchical attributes data are to be aggregated (*group-by set*). The aggregation level has a strong impact on the size of the result returned to the user, and its inappropriate setting may end in either obtaining very coarse, useless information or being flooded by tons of detailed data, which is particularly critical when working with devices with small bandwidth and limited visualization capabilities. For this reason we argue that, in the OLAP domain, users may wish to express their preferences *also* on group-by sets too, for instance by stating that monthly data are preferred to yearly and daily data. On the other hand in the OLTP domain, users typically express preferences *only* on instances. For examples in the hotel reservation domain a user will express a preference stating that “*Hotels with a high level of stars, with low price and close to the city center are preferable*”.

## 2.4 What-if analysis

In order to be able to evaluate beforehand the impact of a strategic or tactical move, decision makers need reliable forecasting systems. What-if analysis satisfies this need by enabling users to simulate and inspect the behavior of a complex system under some given hypotheses, called scenarios. Despite the marketing claims, until now only few tools offered what-if capabilities, and usually they were limited to a specific application. Besides, designing what-if applications requires to understand, simplify, and model business-related phenomena, which may become very difficult in complex enterprises. Unfortunately, no attempt has been made so far to comprehensively address

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 10/18

methodological and modeling issues in this field, while the adoption of naive approaches makes projects more expansive and exposes them to higher risk of failure.

Some challenges arising in this field are: how to seamlessly integrate what-if analysis and OLAP into an extended interaction paradigm; to find an adequate formalism to conceptually express the simulation model on which what-if analysis is centered, so that it can be discussed and agreed upon with the users; to define an appropriate methodology for designing what-if applications, aimed at effectively reducing project failure risks by properly considering user requirements while guaranteeing the model reliability [7].

### 3. Towards the second BI generation

#### 3.1 Pervasive BI

The world of BI is undergoing an important change. While traditionally BI techniques were the realm of a few trained analysts at the top of the enterprise, today they are used from an increasing number of users who exploit BI potentiality to support their daily operations. This mass of new users asks for flexibility not in manipulating data, but rather in accessing data: they require that information can be easily and timely accessed in different contexts, through devices with different computation and visualization capabilities (PCs, palmtops, mobile phones), and with sophisticated and customizable presentations. Besides, access must be selective, depending on the access control and privacy policies adopted in the organization, that could pose some constraints on the type of information each user can see.

Thus, there is a need for advanced techniques for integrated and context-aware management of preferences and security policies in order to achieve personalized and selective access to information in BI platforms, and at the same time guarantee the privacy of data from which information derive. A query context could be defined for instance by the type of device the user is operating, by the user role, by the spatio-temporal location of the user, and by the type of analysis the user is carrying out. Contexts should be associated with preferences and access constraints that, in relationship with a specific user request, will concur in determining the information returned and its presentation format, and consequently the best suited strategy for data processing. The following examples will clarify the expressivity that can be obtained:

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 11/18

- *Show all alerts related to the production chain* (context: role="production manager", device="mobile phone", location="Paris", preferences="recent and local data first"). This query may return an SMS including the three most recent alerts, issued by the factories nearby Paris, regarding the production chain.
- *Show global yearly EU funding for research projects* (context: role="EU officer", device="PC", location="Bruxelles", preferences="computer science best"). This query may return a table summarizing the yearly funding in the computer science area as gathered from all national peers.

### 3.2 Distributed BI

The standard architecture for BI applications is that of a single DW that collects information from (typically local) sources and delivers information that are consumed by the decision makers, mainly using OLAP and reporting applications. In several application contexts, users feel the lack (perceived both at the economic-organizational level and at the technological one) of a framework supporting managers in the identification and use of business information in inter-business collaborative contexts, where several companies organize and coordinate themselves in order to develop common and shared opportunities, respecting their own autonomy and heterogeneity. This need is also felt whenever the time and the budget constraints make it unfeasible to carry out a DW integration project. This is true for instance in the case of company acquisitions where managers need to run queries on DWs of companies belonging to the group. Similarly, for a group of public agencies scattered on the territory, cross-agency analyses could be very useful to obtain overall statistics. We will call *business intelligence networks* (BINs) the consortiums whose goal is the sharing of information at the management level.

Traditional BI system, born to support the decision-making processes of single enterprises, are not capable of operating in inter-company contexts where the (organizational, lexical and semantic) heterogeneity and the autonomy of the individual actors make it impossible to realize a shared and completely integrated BI system. Building a BIN poses several challenges, in particular:

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 12/18

- *Which is the architecture to be adopted:* since one of the requirements is the independence of the participants from the technological and organizational points of view, the architecture must be completely distributed and should be based on a network of peers, each equipped with independent BI systems that expose some information and services. Related issues are how to ensure performance and scalability, and how to ensure security, privacy, and dependability.
- *How to model information:* difference in terminology and semantic of information stored in peers risks to make query results meaningless. Obviously, building mappings manually is unfeasible. A solution could be obtained by coupling the exposed information and services with an ontological model that describes them and enables a mechanism for automatically mapping related concepts in different peers.
- *How to query the network:* exploiting the information stored in the BIN requires a language and a set of algorithms that allow multiple peers to be queried. The querying mechanism should exploit the ontological information stored in the model in order to handle the heterogeneity of the concepts across the BIN and it should be flexible enough to return as much information as possible from different peers. In fact, since different peers can expose only part of the required information, returning only full matches is too restrictive.

The emerging technology related to web services and P2P could be the starting point for a revolutionary approach to BI, where the traditional assumption that DWs contain materialized data is replaced with the novel idea that DWs can contain pointers to web services that guarantee delivery of the data upon request.

### 3.3 Open Source Mapping Systems and Benchmarks

The productivity of each community depends on the ability of its businesses to thrive in the modern knowledge-based global economy. The huge complexity and volume of the data involved in business transactions nowadays has already been recognized. Modern open-source BI tools have recently started to cope with that issue and help businesses explore new opportunities. Modern businesses demand sophisticated data integration tools and techniques, complex analysis solutions and increased strategy planning capabilities.

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 13/18

Existing integration and data exchange solutions can be categorized into two main groups: the *mediated information systems* and the *data warehouses*. The former provide virtual integration of a number of physically distributed heterogeneous sources. This is achieved through a series of software components called *mediators*. Mediators have the knowledge of the schema and the data content of the underlying databases. They provide users and applications with a unified virtual schema. Since mediators have no data, when a query is sent to a mediator, it has to be analyzed decomposed to smaller parts and sent to the individual sources [22]. Then the data returned from the sources is collected by the mediator that combines it and presents it to the user or application that send the query. The advantage of this architecture is that no data replication is necessary and the query results are always up-to-date. In the alternative solution to the mediated architecture, i.e., the warehouse, there is a physical repository in which the integrated data is stored [21]. The advantage of the data warehouse is that queries can be answered on the materialized data, avoiding the need to communicate with the individual sources. The drawback is that the materialized data must be refreshed periodically from the sources, and if not, then the answers to the queries may not be the most up-to-date.

The process of building an information integration or a warehouse system is more similar. First the contents of the data sources are studied and understood. Then a set of queries, commonly referred to as mappings, are designed. Mappings specify the relationship between the schema of the sources and the integrated schema in the mediator or the warehouse. Nowadays, this is mainly a manual task. It requires skilled data administrators that need to spend countless hours trying to understand the data and write the transformations, a laborious and error-prone process. This translates to thousands of euros spent each year by modern businesses and organizations for this task. Intelligent tools that can automate or at least support the data administrators in that task are needed. This has led to the development of the schema mapping tools. A schema mapping tool is a tool that assist a data administrator in defining the mappings through a graphical user interface. It typically displays side-by-side two schemas (referred to as the source and the target schema). The source schema may be the schema of a local source and the target schema the integrated schema in the mediator or a warehouse. Through the interface the data administrator draws lines from the source to the target schema elements representing semantic associations between them. These associations are high level and not expressive enough to describe the full details of the transformations that the data

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 14/18

administrator would have expressed through a transformation query. Thus, the task of the mapping tool is to try to interpret the high level associations and generate the actual transformation queries, saving the data administrator's valuable time and guaranteeing correctness. Of course, due to their fuzzy semantics, there may be multiple interpretations of a given set of correspondences. The role of the data administrator is then to browse these interpretations and guide the tool in selecting the right one.

One of the first schema mapping tools is Clio [20] developed at the IBM Almaden research center. Other similar tools have also been developed, i.e., the Altova MapForce [18] or the Stylus Studio [19]. Unfortunately, none of these tools is open source. Many open source integration solution may provide partial support for the mapping generation task as a side-product, but dedicated open-source mapping generation solutions offering the level of intelligence and functionality of the aforementioned tools, do not exist.

Despite the existing mapping systems, there has been no benchmark developed for comparing and evaluating them. Exception is the STBenchmark [23], a benchmark developed in a collaboration between the University of California Santa-Cruz and the University of Trento. Similar to the motivation of benchmarking relational database management systems, a benchmark for mapping systems is important for assessing their relative merits, which is in turn important to businesses for making the right investment decisions and adopting the tool that best suits their needs. STBenchmark consists of a set of transformation scenarios that have been drawn from the related research literature and from a number of real application scenarios. Through these scenarios one can evaluate the transformation capabilities in terms of functionality of a mapping tool. In addition, STBenchmark offers a schema and data generator software [24]. This allows the generation of complex mapping scenarios that require different multiplexed transformations and can be parameterized to scale to different sizes, allowing the evaluation of the solutions provided by the various mapping tools not only in terms of complex transformations but also in terms of schema and data sizes.

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 15/18

## Conclusions

Thus, in this paper we list and discuss some research topics that are currently being investigated and that, we believe, could suggest new functionalities to be added to OS BI platforms. We envision two main directions: on the one hand, BI platforms need to be consolidated since they still lack in supporting some relevant requirements; on the other hand, part of the new user requirements impose to enlarge the traditional vision of BI since they cannot be fulfilled within the current frameworks. As a consequence, we should begin to work towards a second generation of BI tools that, keeping into account the jumps ahead of both technology and research, will be capable of supporting new types of analysis. BI initiative works in this direction by bringing together researchers as well as practitioners. We believe that BI initiative could be coupled with an Eclipse project aimed at building tools and frameworks for managing data warehouse systems across their lifecycle.

## References

- [1] M. Blaschka, C. Sapia, and G. Höfling. On schema evolution in multidimensional databases. In *Proc. DaWaK*, Italy, 1999.
- [2] B. Bebel, J. Eder, C. Koncilia, T. Morzy, and R. Wrembel. Creation and management of versions in multiversion data warehouse. In *Proc. ACM SAC*, Cyprus, 2004.
- [3] J. Chomicki. Preference formulas in relational queries. *ACM Trans. on Database Systems*, 28(4), 2003.
- [4] R. Kimball, L. Reeves, M. Ross, and W. Thornthwaite. *The data warehouse lifecycle toolkit*. John Wiley & Sons, 1998.
- [5] M. Golfarelli D. Maio, and S. Rizzi. The Dimensional Fact Model: A conceptual model for data warehouses. *Int. J. Cooperative Inf. Syst.* 7(2-3), 2002.
- [6] M. Golfarelli, J. Lechtenböcker, S. Rizzi, and G. Vossen. Schema versioning in data warehouses: Enabling cross-version querying via schema augmentation. *Data & Knowledge Engineering*, 59(2), 2006.

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 16/18

- [7] M. Golfarelli, S. Rizzi, and A. Proli. Designing what-if analysis: Towards a methodology. In *Proc. DOLAP*, Arlington, USA, 2006.
- [8] L. Gonzales. Seeking Spatial Intelligence. *Intelligent Enterprise Magazine*, 2(3), 2000.
- [9] W. Kießling. Foundations of preferences in database systems. In *Proc. VLDB*, China, 2002.
- [10] J. Hammer, M. Schneider, and T. Sellis. *Data Warehousing at the Crossroads*. Dagstuhl Seminar 04321. Germany, 2004.
- [11] A. Maniatis, P. Vassiliadis, S. Skiadopoulos, and Y. Vassiliou. Advanced visualization for OLAP. In *Proc. DOLAP*, USA, 2003.
- [12] Y. Sismanis, A. Deligiannakis, N. Roussopoulos, and Y. Kotidis. Dwarf: Shrinking the PetaCube. In *Proc. SIGMOD*, USA, 2002.
- [13] OLAP Report. <http://www.olapreport.com/market.htm>.
- [14] S. Rizzi. OLAP preferences: A research agenda. In *Proc. DOLAP*, Portugal, 2007.
- [15] SAP. Multi-dimensional modeling with BW. Technical report, SAP America Inc. and SAP AG, 2000.
- [16] C. Thomsen and T.B. Pedersen. A survey of open source tools for business intelligence. In *Proc. DaWaK*, Denmark, 2005.
- [17] A. Vaisman, A. Mendelzon, W. Ruaro, and S. Cymerman. Supporting dimension updates in an OLAP server. In *Proc. CAiSE*, Canada, 2002.
- [18] Altova MapForce. <http://www.altova.com/products/mapforce/>
- [19] Stylus Studio. <http://www.stylusstudio.com>
- [20] L. Popa, M. Hernandez, Y. Velegrakis, R. J. Miller and R. Fagin. Mapping XML and Relational Schemas with Clio. Demonstration in IEEE International Conference in Data Engineering (ICDE), Feb. 2002.
- [21] Ralph Kimball: *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. John Wiley 1996.
- [22] Maurizio Lenzerini: *Data Integration: A Theoretical Perspective*. In *PODS*, pages 233-246, 2002.

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 17/18

- [23] STBenchmark: A benchmark for Schema Mapping tools. <http://www.stbenchmark.org>
- [24] B. Alexe, W. Chiew Tan, Y. Velegrakis: STBenchmark: Towards a Benchmark for Mapping Systems. In VLDB 230-244, 2008

OW2 BI Initiative deliverables	OS BI: Open Research Issues	December 22 <sup>nd</sup> , 2008
Author: DEIS, University of Bologna DISI, University of Trento	Revision: Engineering, University of Milan	Page: 18/18