Intelligent Data Sources and Integrated Data Repository as a Foundation for Business Intelligence Analysis

Damir Pintar, Mihaela Vranić, Zoran Skočir

Faculty of Electrical Engineering and Computing,
University of Zagreb, Croatia
E-mail: damir.pintar@fer.hr ; mihaela.vranic@fer.hr; zoran.skocir@fer.hr

Abstract: **Data mining and data analysis in general demonstrate high dependency on data quality. Gathering the right data of high enough quality takes most of researcher's time and often demonstrates need for some additional data to be parsed. In order to eliminate or at least reduce required effort for this first phase of every analysis, authors of this paper present the idea of Integrated Data Repository and Intelligent Data Source. Concepts of those components are presented and approach to their development is suggested together with the high-level view of the system architecture. Finally, an experimental implementation is described.**

1. INTRODUCTION

The efficiency of business information analysis methods, such as data warehousing and data mining processes, is largely influenced by the quality of available data. Extracting, transformation and loading (ETL) of data is often the most time-consuming and arduous task without which further steps of the analysis are impossible. The ETL process prepares the data by gathering it, cleaning it and transforming it in the appropriate form. Data first has to be collected from various distributed sources, different formats of the data must then be consolidated, mistakes removed, missing data should be dealt with in appropriate manner and finally the prepared information should be stored in such a way the analysis applications can easily access it and use it accordingly.

With today's technologies, when the capacities of different media allow storage of vast amounts of information and establishing communication between different systems is much easier due to the development of standardized communication protocols and data formatting techniques, gathering large amounts of data from various systems must not present a heroic venture it once was. If the focus of business analysis improvement is changed from *analyzing past results better* to providing means for *improved collection of higher quality data in the future* then some valuable insights and conclusions can be made - the result of this kind of shift in perspective is discussed further in the paper.

One must not dismiss the need for up-to-date reports in business intelligence systems on the other hand. Having the right information in right moment may be crucial for company's future. As stated in reports of prominent advisory companies, global providers of market intelligence, boost in number of users of business intelligence is expected [1], [2], [3], [4]. Those users come from various departments and levels of organizations. They all have to see up to date information, and the information that fit their needs, view of business, responsibilities and restrictions. All sorts of tools that could represent data in different fashions are at the software market. The main problem is to have the right data and that data should be of high enough quality.

This paper gives a presentation of architecture for organized data gathering from distributed sources based on XML, Web Services technologies and the concept of an Intelligent Data Source. Taken into account that an effective conceptual module for the ETL process is of grand importance [5], the main idea of this paper is to avoid the tenuousness of the ETL process by *collecting the operational data at the point of its creation.*

The paper is structured as follows. In section 2 the concept of Integrated Data Repository is presented accompanied with introduction to process of structuring it. Section 3 firstly gives insight in conception of Intelligent Data Source after which guideline for implementation is given. Section 4 presents overall architecture of suggested data gathering system. Section 5 presents an experimental implementation of proposed architecture. Finally, in section 6 the conclusion is given.

2. INTEGRATED DATA REPOSITORY

2.1 The system environment

The initial assumption about the environment was based on the situation found in the large number of distributed information systems today, especially the ones which came into existence by merging different systems in the course of time. The system environment taken into account is supposed to be built of a number of distributed functional hardware and software modules, each self-contained and performing its own function while communicating with other modules by previously assigned protocols. The type, quantity and exact contents of data shared between modules are governed by the specific business function the modules perform as a whole. Each module gathers and/or generates a certain amount of data according to its operating needs, and those operating needs dictate how the operation gets handled inside the module. It is assumed that no integrated data store exists when the system is seen as a whole, or at least not as such to include the entire information system data gathered throughout the system during its daily operation.

These kinds of information systems are often found in large corporations or various environments which are not directly business-oriented (like for example academic [6] or government environments). Profit companies confront similar problems when acquisitions and joins are taking place. When the need for various business intelligence analyses arises in these kinds of environments, this inevitably results in an expensive and resource-intensive ETL process which has already been discussed in the Introduction.

2.2 Constructing the Integrated Data Repository

The first order of business for implementing the architecture discussed in this paper is the construction of an integrated data repository – a "container" of sorts which would serve as the main source of information for the analysis expected to be performed in the future. This repository should exist in parallel with the operational systems or functional modules as mentioned in the previous paragraph. The repository should store the required information but at the same time not obstruct the daily operations and business functions of the modules in any way.

In the past, focus on different aspects of business was changed in the most companies. In the beginning companies kept track about their primary course of business. Afterwards, customers are being slowly recognized as very important subjects in business and additional data about them is collected and analyzed. In the next step the employees themselves became a relevant factor for business success so much more data about their performance, education and other aspects was collected. Today every bit of information could be significant to keep track about. It is difficult to foresee in advance the exact importance of each piece of information - even some data that seems to be of no value at one particular moment could later turn to be appreciable. So the conclusion would be to store as much data as possible in the Integrated Data Repository. However, dependencies and relationships between data from different modules should be investigated and documented possibly in form of structured semantic ontology.

The first step should be building a formal description of an information system – a conceptual model which will represent real-world objects the analysis will be interested in. A semantic model would be preferable ([7],[8]) since it would help with future potential terminology issues which often turn up when data of similar type is gathered at different sources. Nevertheless, the optimal model should be chosen to adapt to the particularities of an actual implementation - an E-R model could be sufficient depending on the environment. In any case, this model will serve as a Meta Data model for the future repository, similary to [9].

Conceptual model of the real-world objects of interest should then be mapped to a set of tables (which will effectively serve as the container for the data). These tables should be governed by additional business logic above – this logic would be responsible for receiving and storing the data as well as certain book-keeping tasks. Also, data repository should contain a register of known data sources and ascertain how to deal with them and the data received. This all functions as an integrated data repository which will gather the data from various sources and store it in an organized fashion.

The whole process of building an Integrated Data Repository and its subsequent usage is presented in Figure 1. The process begins by studying existing information systems and needs for different information by employees throughout organization. Here, different definitions of data must be reconciled and granularity of data that will be stored must be determined. Also expansion of requirements must not be overlooked and must be enabled as much as possible. Connections between different data classes (entities) must be established, similarly to way presented in [10]. These semantic structures are further referenced when implementing each Intelligent Data Source which will send certain data to the Repository. The process of establishing right semantic structure is iterative and time consuming but necessary for collecting quality data. After completion of this task, relational database can be developed. Finally, in the working phase, Repository must enable parallel and undisturbed data load, system maintenance and access to the stored data. System maintenance, beside usual activities like enabling access to new users, also includes adding new attributes to existing objects but also adding new objects of interest to the database structure. Relation system enables those operations pretty good. Mentioned operations of changes in relational system must be also reflected in base semantic structure.

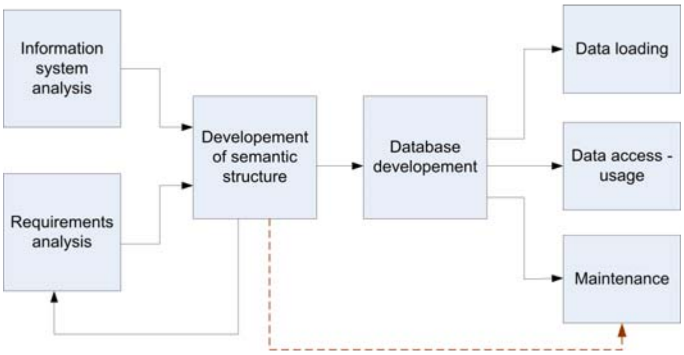

Figure 1 –Building of Integrated Data Repository

Gathered data, or at least certain portions of it, can finally be used for different analyses – either for business intelligence in broader sense or purely for data mining purposes. It is also important to emphasize due to security concerns - nowadays database systems enable fine tuning of responsibilities for employees at different positions in organization - so they can be expected to be able to access only certain pieces of data for which they are authorized.

3. THE INTELLIGENT DATA SOURCE

3.1 The concept of Intelligent Data Source

We have found the term "intelligent" to be fitting for the entity responsible for gathering and sending the data from one specific business module to the repository because of the following reasons:

- IDS (Intelligent Data Source) knows *which information* it should collect
- It knows *when*, *where*, and *how* to collect it
- It knows in which way the gathered information should be *transformed* so it would fit the formats dictated by the repository
- It knows *where and in what manner* it should send the information so it would be appropriately stored.

Intelligent Data Source entity should be placed in (or closely coupled with) the location where the information is generated. It should ideally work in conjunction with the functional module it resides close with but without obstructing the operational business cycle in any way. For example, if the IDS were assigned to collecting information about received purchase order in a certain Commerce module, it should gather all the required data from the received purchase orders but at the same time shouldn't interfere in any way with the actual ordering business process.

3.1 Implementing the Intelligent Data Source

The exact implementation of the IDS would be largely dictated by the environment where the information it should collect gets generated. For example, in the environment where the information gets stored in certain tables in a relation database, that particular IDS could be partly implemented as a database trigger waiting for certain conditions (e.g. entering of new purchase order data) with appropriate following actions.

Certain qualities should be shared regardless of the exact IDS implementation. Every IDS should contain:

1) Logic that transforms the information it is assigned to gather to the previously defined form (preferably XML)
2) Logic that sends the data to the repository via previously established protocols (the Web Services technology is suggested).

Therefore, Intelligent Data Source is basically an intelligent data collection facility – it realizes that the "interesting" data has appeared, it collects it and then transforms and sends it. This is largely what the typical ETL process is about – collecting, transforming and storing the data in certain way. The major difference is that ETL applications work with the data previously collected in the certain time periods, while IDS is built to work with future data, so it can be finely tuned right at the start to ensure that data will be of maximum quality before the analysis has even begun its initial phases.

4. ARCHITECTURE

After the design and construction of the Integrated Data Repository as well as establishing a number of Intelligent Data Sources is finished, there remains the task of integrating everything into a functional system. Figure 2 shows the perceived architecture described further on.

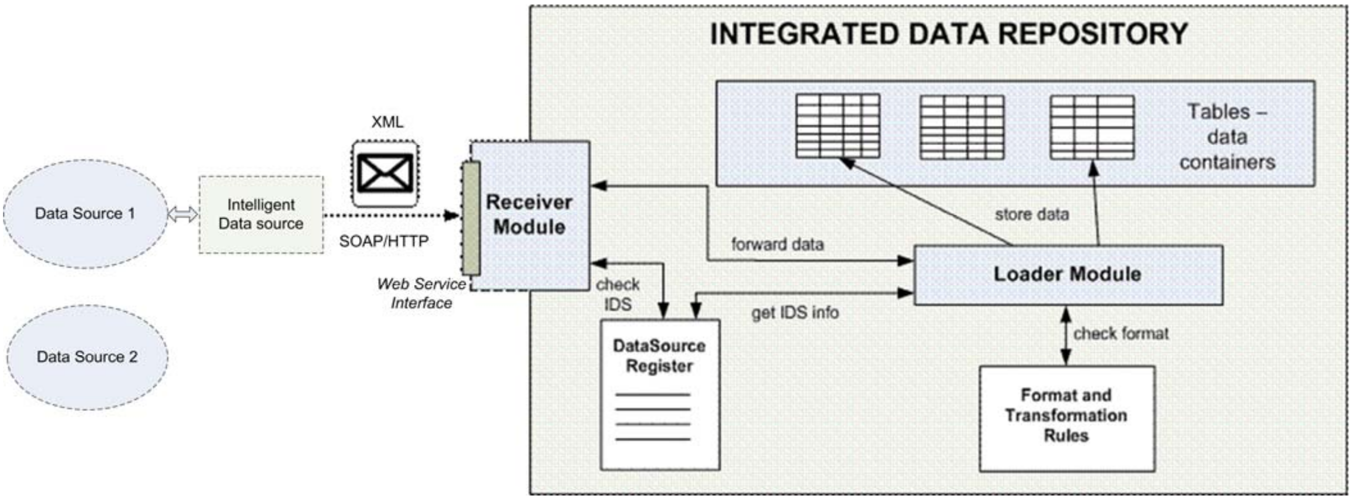

Figure 2 –Architecture of Integrated Data Repository and communication with Intelligent Data Source

The proposed protocols for communication between the Data Sources and the repository is Web Service technology – using HTTP and SOAP as communication protocols with XML data as payload.

The repository has a Web Service module called the Receiver. This module is reachable on certain URL and serves as the main entrance point in the repository. Every Data Source that uses this repository must be previously registered – a list of registered Data Sources is kept in a table in the repository itself. After the Receiver receives the data it first checks the validity of the sender (by following an appropriate security protocol) and then forwards the data received to the Loader module.

Loader module analyzes the data and checks whether it conforms to the established standards. If the data is deemed valid it is stored in the appropriate tables. If not, an exception is raised and certain measures must be undertaken to look into what the exact problems are and how they should be worked out. Also, every bit of information is coupled with the data about the source of information and the timestamp, which would facilitate future analysis ventures.

Data about Intelligent Data Sources stored in Repository is twofold: firstly, some attributes that are assigned to specific Intelligent Data Source when developed and registered concerning data it gathers and information system module that provides data for that specific Source. This data is used for authorization purposes and basic formatting guidelines of the data collected. Other data about it is created dynamically and reflects characteristic of specific Source concerning dynamic of its deliveries and data quality delivered.

Ideally, the system should be fully automatic with minimal human interference. The repository would keep track of all the information from the environment deemed important during the course of daily business operations. If new sources of information are identified, after subsequent IDS implementation they should be registered with the system after which they can start sending in the information.

5. IMPLEMENTATION

As a part of an ongoing project, the authors of this paper are involved in a design of an infrastructure which will be used to track the software and hardware resources of a specific organization. Information about the current situation has to be gathered and future changes have to be closely monitored. Gathered data will be used for various analyses with the main goal being to improve the cost-effectiveness of the entire system.

This scenario was found to be an adequate testing ground for the architecture proposed in this article. A data repository had to be defined to store all the information in, a way for collecting the initial data had to be procured and finally a long-standing mechanism for automatic (or semi-automatic) gathering of future data had to be implemented.

5.1 The implementation of data store

As it was mentioned in Section 2., the design of the data store was started with an ER model. The ER model was then used to build a relational database. Also, a semantic model which corresponded with the ER model was developed in parallel.

This semantic model was used to build XML schemas for the actual documents which will contain gathered data. Initially, two types of documents were developed – *HardwareInfo.xsd* and *SoftwareInfo.xsd* which respectively held information about hardware and software resources on one particular computer system.

Each schema then got a corresponding XSLT file whose purpose was to transform the received XML document in a series of SQL statements which would be used to enter the information in the database. To increase flexibility, a *properties* file was also procured which contained the database connection details so they wouldn't have to be hard-coded later.

The XML schemas, XSLT documents and the *properties* file represented the “Format and Transformation rules” module shown on the Figure 2. The Loader module was then implemented as a Java application module (*LoadingModule.class)* whose purpose was reading and validating the received XML document and loading the information in the database with the help of mentioned documents.

For simplicity's sake due to mentioned time constraints, the Receiver module was designed in the lightest way possible – as a Web service module whose role was providing an entrance point for XML documents with only provisional security mechanisms. The Datasource Register was not used in this scenario.

All these elements constituted the proposed Integrated Data Store. The "container" for the information was set, and the next step was providing means for gathering and sending the information.

5.2 The implementation of Intelligent Data Source

Gathering the information was expected to be conducted in the semi-automatic way – some information about hardware and software resources could be provided automatically, while some had to be entered by the users manually (mostly the opinions and views of users about the importance and usefulness of certain applications).

It was decided that the most elegant solution was to create JSP pages populated with fields which correspond to elements of XML documents mentioned in Section 5.1. Some of those fields would be filled automatically while some required the user to enter the important information manually. After all the necessary fields are filled, clicking on the "Submit" button bundles all the information in an XML document, timestamps it and then sends it to the Web service of the Receiver module. This corresponds to the concept of Intelligent Data Source.

To cope with the fact that the system has to be able to collect future data, the current mechanic is to remind user (or users) of certain computer system to re-visit the JSP pages on bi-monthly basis. A more efficient system would remove the user from the loop and automatically decide whether re-entering the information for a certain computer system is required and whether the user has to be contacted about it. This is left for future work after testing the effectiveness and functionality of the current system.

6. CONCLUSION

Today, access to the right and exact information at right time is crucial for business success. Also data mining analysis can provide useful information to companies. The main obstacle here is data quality and information systems that consist of different modules that are not in sink. Getting to the right data with acceptable definitions could present real issue.

The solution presented in this paper gives an alternative to the resource-intensive ETL process which is an obligatory step in the business analysis process. By establishing an infrastructure for collecting relevant information *in advance* and storing it in adequate form, this process can be facilitated and speeded up drastically.

The concept of Integrated Data Store and Intelligent Data Sources is relatively painless to implement – it uses well known open technologies and doesn't require a major financial investment. Furthermore, this type of architecture supports – and can even benefit from – the evolutionary approach to building the system. A relatively small data store built on a simple semantic model and connected to a few data sources could have its experimental but also actual business use. As the needs and requirements scale up, the system can follow suite by expanding the model it is based on, adding new tables and transformation rules, registering new data sources and finally storing additional data.

This paper presents a high-level solution of the system – one implementation was shown but the particularities of the specific implementation and dependence on certain technologies and protocols should be left open. Further work will mostly include investigation of the ways of using the semantic model most beneficially in the initial design phases, but especially at later time when automatic data transformation can become a big issue.

REFERENCES

[1] Dan Vesset: “The next wave of business intelligence”, SAScom magazine, pp. 5-6, first quarter 2007.

[2] Jim Goodnight, Keith Collins, Mikael Hagström, Jim Davis: “Future of business intelligence”, SAScom magazine, pp. 16-22, first quarter 2007.

[3] IDC reports: http://www.idc.com/

[4] Gartner reports: http://www.gartner.com/

[5] P. Vassiliadis, A. Simitsis, S. Skiadopoulos: “Conceptual modeling for ETL processes”, Proceedings of the 5th ACM international Workshop on Data Warehousing and OLAP (McLean, Virginia, USA, November 08 - 08, 2002). DOLAP '02. ACM Press, New York, NY, 14-21.

[6] M. Vranić, D. Pintar, Z. Skočir: “The use of data mining in education environment”, Proceedings of the 9th International Conference on Telecommunications, ConTEL 2007, Zagreb, Croatia, pp. 243-250

[7] D. Juric, Z. Skočir: “Building OWL ontologies by analyzing relational database schema concepts and WordNet semantic relations”, Proceedings of the 9th International Conference on Telecommunications, ConTEL 2007, Zagreb, Croatia, pp. 235-242

[8] D. Skoutas, A. Simitsis: “Designing ETL processes using semantic web technologies”, Proceedings of the 9th ACM international Workshop on Data Warehousing and OLAP (Arlington, Virginia, USA, November 10 - 10, 2006). DOLAP '06. ACM Press, New York, NY, 67-74.

[9] G. Lao, Y. Tang: “The application of data warehousing in e-business environment and case study”, Proceedings of the 7th international Conference on Electronic Commerce (Xi'an, China, August 15 - 17, 2005). ICEC '05, vol. 113. ACM Press, New York, NY, 815-817.

[10] M. Banek, B. Vrdoljak, A Min Tjoa: “Using Ontologies for Measuring Semantic Similarity in Data Warehouse Schema Matching Process”, Proceedings of the 9th International Conference on Telecommunications, ConTEL 2007, Zagreb, Croatia, pp. 227-234