Computer Science • 15 (2) 2014 http://dx.doi.org/10.7494/csci.2014.15.2.109

Marek Macura

INTEGRATION OF DATA FROM HETEROGENEOUS SOURCES USING ETL TECHNOLOGY

Abstract *Data integration is a crucial issue in the environments of heterogeneous data sources. At present, the afore-mentioned heterogeneity is becoming widespread. Based on various data sources, if we want to gain useful information and knowledge, we must solve data integration problems in order to apply appropriate analytical methods to comprehensive and uniform data. Such activity is known as knowledge discovery from the data process. Therefore, approaches to the data integration problem are very interesting and bring us closer to the "age of information". This paper presents an architecture which implements knowledge discovery from the data process. The solution combines ETL technology and a wrapper layer known from mediated systems. It also provides semantic integration through connection mechanism between data elements. The solution allows for integration of any data sources and implementation of analytical methods in one environment. The proposed environment is verified by applying it to data sources in the foundry industry.*

Keywords data integration, integration approaches, ETL technology, knowledge discovery from data, business intelligence

1. Introduction

The current growth of IT results in the development of various solutions, both in terms of hardware and software. Along with this growth, the solutions have become more common, and computerization of society has been increasing. There are many vendors, and their products are often not compatible with each other. Institutions and companies are forced to adapt solutions from different vendors to support their activities in a wider range. Current hardware capabilities allow us to collect and process vast amounts of an institution's operational data.

The processing of operational data should allow institutions to obtain important information that, properly interpreted, can also provide knowledge. Knowledge discovery is the most desirable end-product of computing. Finding new phenomena, or enhancing our knowledge about them, has a greater long-range value than i.e., optimization of the production process. It is not surprising that this is also one of the most difficult computing challenges. As mentioned, current technological progress permits the storage and access of large amounts of data. However, to paraphrase Galileo, "the accumulation of data is still not science". The true value is not in storing data, but rather in our ability to extract useful reports and to find interesting trends and correlations through the use of statistical analysis and inference to support decisions and policies made by scientists and businesses [15].

Gathered data often comes from heterogeneous (different) sources; therefore, integration activities are needed and very important. In a business context, integration activities are commonly referred to as Enterprise Integration. This means the ability to integrate information and functionalities of different IT systems. EI includes Enterprise Information Integration, which refers to integration of data, and Enterprise Application Integration, relating to integration on the level of application logic.

This paper refers to the integration of data, used in order to connect and provide unified access to it. Due to the heterogeneity of data sources, such activities are difficult to solve. The heterogeneity of data sources refers to differences in their architectures. Here, we can distinguish different access methods and, in particular, heterogeneity at the data level (which appears in two forms: syntactic and semantic). To solve the data integration problem, we must develop a data integration system that reduces disparity between the current ability to gather, manage, and analyze data. We can say that such systems bring us closer to the "age of information". Therefore, such approaches and architectures are constantly being investigated and have been of particular interest in the science and business context.

The goal of this paper is to present a proposed architecture that implements knowledge discovery from the data process and its verification result. Because data integration is the basis for knowledge discovery from the data process, the proposed architecture is most important in a data integration system. The most mature and applied approach to data integration is ETL technology. However, it has some deficiencies. ETL does not support types of data sources such as novel kinds of data (websites, spatial, or biomedical). ETL tools are only used for data integration and do not provide an analytical environment. The proposed solution allows us to integrate any type of data, and provides the analytical environment as a single component. The solution does not automate knowledge discovery from the data process, but supports user activities requiring the involvement of intelligence. The created prototype refers to integration of data related to the foundry industry. This paper is structured as follows: Sect. 2 describes the problem of data integration. Sect. 3 presents principal approaches to data integration. Sect. 4 describes knowledge discovery from the data process and its connection with data integration. In Sect. 5, data warehouses and ETL technology are discussed. In Sect. 6, the proposed solution is presented. Sect. 7 contains a description of the practical application. Sect. 8 contains a description of the verification result. Sect. 9 concludes the paper.

2. Data integration problem

Data integration is an area of research that addresses a pervasive challenge faced in applications that need to query across multiple autonomous and heterogeneous data sources [14]. Data integration is crucial in large enterprises, for large-scale scientific projects, for better cooperation among government agencies, and in offering good search quality across huge amounts of data on the World Wide Web [5].

Data integration *is the problem of combining data residing at different sources, and providing the user with a unified view of these data* [16].

Data integration addresses problems related to the provision of interoperability to information systems by the resolution of heterogeneity between systems on the level of data. The problem of data integration can be decomposed into the following subproblems [14]:

- Structural integration. It refers to the resolution of structural heterogeneity; for instance, the heterogeneity of data models, query and data access languages, protocols, and hardware platforms.
- Semantic integration. It refers to the resolution of semantic mismatch between schemata. A mismatch of concepts appearing in such schemata may be due to a number of reasons. For instance, different schemas may represent the same information in different ways.

The major issues that make integrating data difficult include [19]:

- Heterogeneity of the data sources. Each data source might have a different data model. The representation of data of similar semantics might be quite different in each data source. Moreover, they may contain conflicting data. In addition, heterogeneity may also occur at lower levels, including access methods, underlying operating systems, etc.
- Autonomy of the data sources. Data sources are independent elements that are not designed for a data integration systems. They cannot be forced to act in certain ways. As a natural consequence of this, they can also change their data or functionality unannounced.

- Query correctness and performance. Queries to integrated system are usually formulated according to the unified model of the system. The issues of proper processing and its performance are very important.
- Distribution. It refers to the physical distribution of data sources. The appropriate system architecture should take into account the possible latency to communicate with data sources.

To solve the data integration problem, usually a data-integration system is designed. The main contribution of this system is that users can focus on specifying what data they want rather than describing how to obtain it [19, 5].

A data integration system is basically an information system. The data sources must be integrated as they are without making any changes on their design and operation. Also, like all information systems, there is an application domain that a system has to model. It is obvious that underlying data sources determine this application domain. A data-integration system also has to provide query functionality, and it depends on query capabilities of underlying data sources [19].

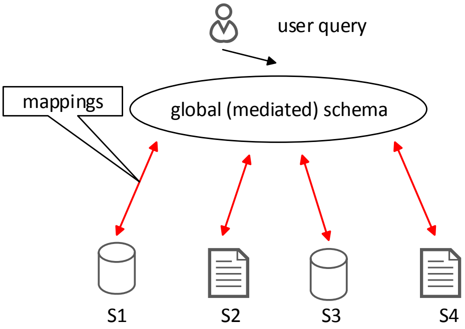


Figure 1. Higher-level abstraction of data integration [4].

On the higher-level abstraction, a data integration system could be considered as three elements. The first one is a global schema (or mediated schema), which provides a reconciled, integrated view of the underlying sources and query-interface to access data. The second one is a set of source schemas, each schema describes data in underlying source. The last one is a set of semantic mappings between the global schema and sources. It is shown in Figure 1.

There exists a theory [16] associated with the previously presented idea of a data integration system that is a subset of database theory and presents basic concepts of integration in the form of first-order logic. This theory contains information about the possibilities and difficulties of data integration. Represented abstract concepts are general enough to associate all kinds of data integration systems with them.

3. State of the art

Data integration from several sources is performed based on the needs of one or more end users. These requirements are represented by a view called the integrated view, which basically is a global schema. There are several research and commercial projects in data integration whereby each one proposes an integration approach. Two major approaches have been proposed to this problem, both of which are end-points along a broad continuum of possible implementations [9], and these are [19]:

- Virtual View Approach. Data is accessed from the sources on-demand when a user submits a query to the system. This is also called a lazy approach to data integration.
- Materialized View/Warehousing Approach. In this case, some filtered information from data sources is materialized in a repository (warehouse) and can be queried later by users. This is also called an eager approach to data integration.

Both approaches take a set of pre-existing data sources related to a particular domain, and they provide a single unified (mediated) schema for that domain [9]. There also exists a hybrid data integration approach. In data integration systems, there is a trade-off between query response time and data freshness. Fully materialized and fully virtual approaches favor one of these objectives. Hybrid approaches have been developed to create optimal and more-flexible integration systems. They try to give the possibility of materializing some underlying data and querying others in a virtual manner [11, 10, 7].

The main representative of the virtual view approach is an architecture of mediated systems. These systems integrate any heterogeneous data sources by providing a virtual view of all data. The system presents one global schema (called mediated schema), and users pose their queries in terms of it. Then, the user's query is decomposed into sub-queries to individual sources based on their descriptions. In the next step, these sub-queries are sent to the wrappers of individual sources, which will execute them over local models and schemas of sources. Then, a mediator receives answers from wrappers, combines them into one answer (single representation), and sends it to the user. The mediation approach introduced in paper [24] can be characterized as an approach that provides an intermediate infrastructure (middleware) that simplifies query evaluation of several autonomous and heterogeneous local sources. This middleware has to overcome the problems of heterogeneity and concurrency, and it must provide query optimization. Figure 2 depicts the typical architecture of a mediation system. In this architecture, a central mediator is defined as a virtual view and contains a global integrated schema of underlying sources. For each source, a wrapper (translator) is developed that contains the information of mapping between the global schema and the schemas of the corresponding sources. A mediated system uses a common data model and a common query language at the mediation level which may be different from that of some or all data sources [11].

In a materialized-view approach, data from various sources is integrated by providing a unified view of this data, like in a virtual view approach. But here, this

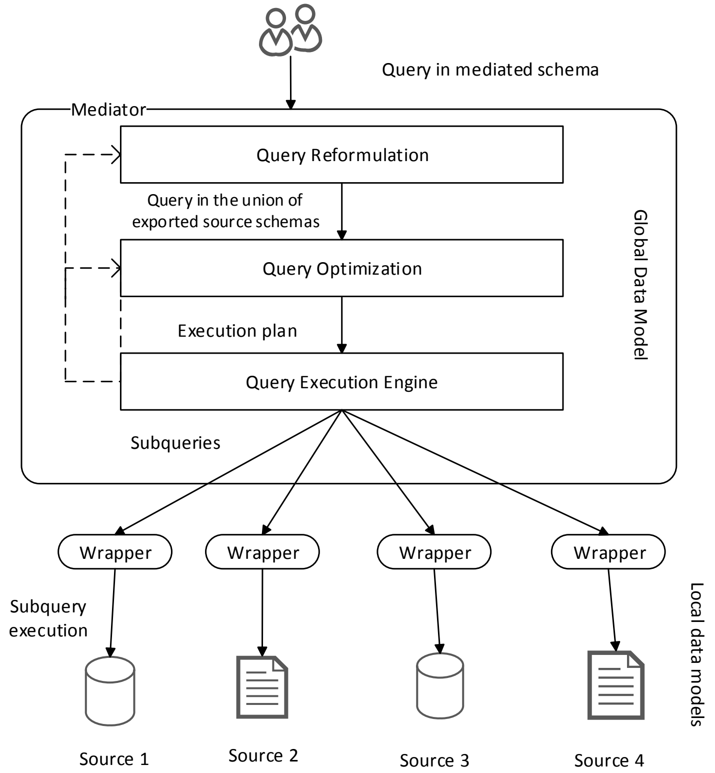


Figure 2. Mediated systems architecture [17].

filtered data is actually stored in a single repository. The original (and still predominant) approach to data integration in the enterprise is a centralized database called a data warehouse. In this model, all necessary data is translated into a target schema and copied into a single DBMS, which periodically gets refreshed. In addition, the transformations that are used to load the data into the warehouse are typically carried out by pipelines of procedural code. The wide variety of tools used to populate data and maintain a data warehouse over time are generically referred to as ETL, or the extract/transform/load tool. The ETL tool addresses a wide variety of tasks: import filters, data transformations, deduplication, profiling, and quality management. Data warehouses serve the natural role of archival and decision support in business. More generally, the warehouse (as a consistent "global snapshot" of an enterprise's data) can be used to perform the so-called decision support or online analytic processing (OLAP) queries [4]. The basic architecture of data warehouses from data integration perspective is shown in Figure 3.

In some studies [23, 1, 3, 2], data integration using materialized views is known as a data warehouse. A materialized view is a database object that contains the

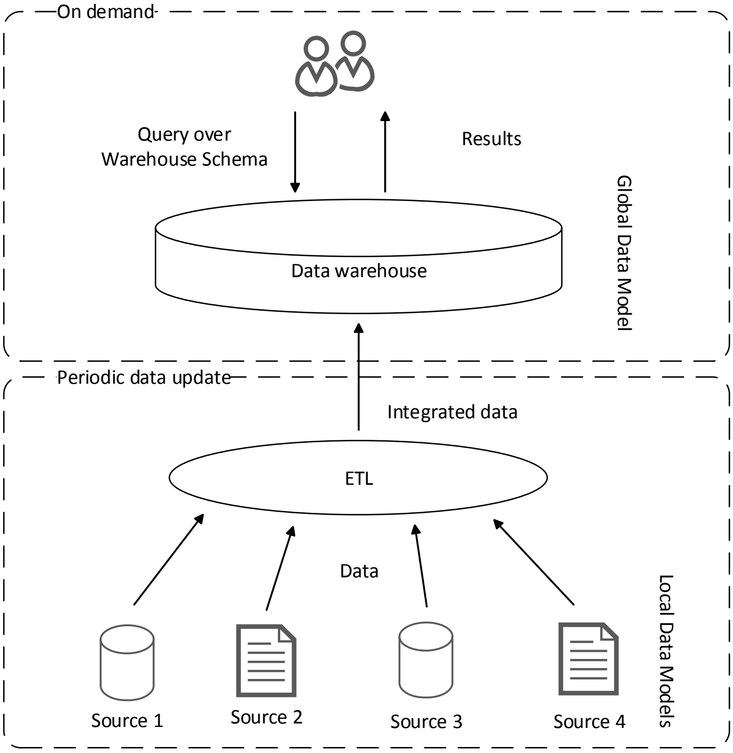


Figure 3. Data warehouse architecture [9].

result of a query. Thus, a materialized view (like a cache) is a copy of data that can be accessed quickly [11]. This solution leverages mediated architecture known from a virtual approach. It is used for theoretical considerations and refers mainly to the databases. Such an approach is defined as query-driven, where the main factor of interest are queries and the efficient processing of them. This is not an equivalent solution to the data warehouse. By using ETL tools, data warehouses employ an update-driven approach, where the main factors of interest are the processes of data update and integration [6].

The virtual view approach is preferable in the following cases:

- data in the sources can change quickly,
- queries are unspecified,
- queries are related to a large amount of data from multiple sources.

In this approach, performance issues may occur related to the processing of queries, especially when many of them are performed, the efficiency of data sources is low, data sources are periodically not available, and when there is a need to perform specific processes for transformation, filtering, and combining data. This approach cannot be used when the sources do not allow for ad-hoc queries. Whereas, a materialized view approach is preferable in the following cases:

- users have specific requirements with respect to data;
- efficient query processing is required and less attention is focused on data update;
- users require access to private copies of data for modification, annotation, summary or analysis;
- when we want store data that is not supported by local data sources (i.e., historical data).

Hybrid solutions are generally considered as a way to improve the efficiency of mediation systems and, therefore, the virtual approach. Some queries can be materialized in the form of a new source. But this is not a solution widely described in the literature, such as the two presented earlier.

4. Knowledge discovery from data

The data stored by different organizations are often heterogeneous in origin, content, and representation, and concern different areas. This data originates partly from internal transactions of an administrative, logistical, and commercial nature, and partly from external sources. They need to be processed by means of appropriate extraction tools and analytical methods capable of transforming them into information and knowledge that can be subsequently used by decision makers. The difference between data, information, and knowledge can be better understood through the following remarks [22]:

Data. Generally, data represent a structured codification of single primary entities as well as transactions involving two or more primary entities.

Information. Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain. For example, to the sales manager of a retail company, the proportion of sales receipts in the amount of over 100$ per week represents a meaningful piece of information.

Knowledge. Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. Therefore, we can think of knowledge as consisting of information put to work into a specific domain, enhanced by the experience and competence of decision makers in tackling and solving complex problems.

We can obtain knowledge in a process called knowledge discovery from data (KDD), which is presented on Figure 4 and consists of the following elements:

- Data integration (also cleansing),
- Data selection (where data relevant to the analytical task is retrieved),
- Data transformation (where data is transformed and consolidated into the appropriate form for mining by performing summary or aggregation operations),

- Analytical technologies (i.e., reports, OLAP, pattern evaluation, data mining, and visualization),
- Information interpretation (performed in order to obtain knowledge).

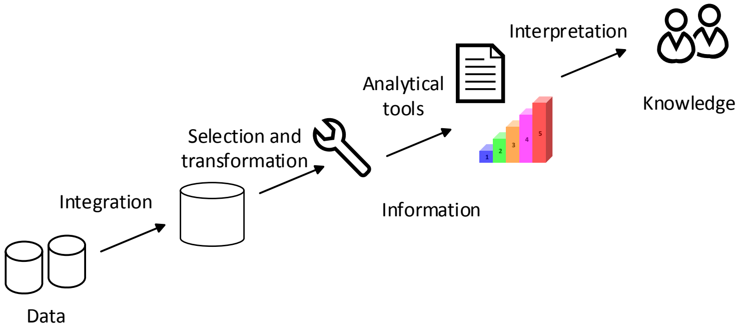


Figure 4. Knowledge discovery from the data process [6].

The basis of knowledge discovery from the data process is data integration. Data integration systems are also information systems, so users can gain useful information and knowledge. The practical solution that uses the described process is Business Intelligence. It is used to understand the capabilities available in the firm; the state-of-the-art, trends, and future directions in the markets, the technologies, and the regulatory environment in which the firm competes; and the actions of competitors and the implications of these actions. Business intelligence is a natural outgrowth of a series of previous systems designed to support decision making. The emergence of the data warehouse as a repository, the advances in data cleansing that lead to a single truth, the greater capabilities of hardware and software, and the boom of Internet technologies that provided the prevalent user interface all combine to create a richer business intelligence environment than was previously available [18]. Business intelligence may be defined as a set of mathematical models and analysis methodologies that systematically exploit the available data to retrieve information and knowledge useful in supporting complex decision-making processes [22].

5. ETL technology and data warehouse

One of the most influential advocates of the data warehouse is the American computer scientist Bill Inmon, who gained the title of "father of the data warehouse" due to actively promoting this concept. Referring to the definition created by him: **Data warehouse** *is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions.*

Thus, in the cited definition data are [8]:

- Subject-oriented. It means that data in the data warehouse is organized so that all of the data elements related to the same real-world event or object are linked together. Typical subject areas are i.e., customer, product, order. Each type of company has its own unique set of subjects.

- Integrated. Data is fed from multiple, disparate sources into the data warehouse. As the data is fed, it is converted, reformatted, resequenced, summarized, and so forth. The result is that data - once it resides in the data warehouse - has a single physical corporate image. Of all the aspects of a data warehouse, integration is most important.
- Time-varying. The changes to the data in the data warehouse are tracked and recorded so that reports can be produced showing changes over time. Different environments have different time horizons associated. Time variance implies that every unit of data in the data warehouse is accurate at some moment in time.
- Non-volatile. Data in the data warehouse is never over-written or deleted - once committed, the data is static, read-only, and retailed for future reporting. Data is loaded and accessed, but it is not updated. When subsequent changes occur, a new snapshot record is written.

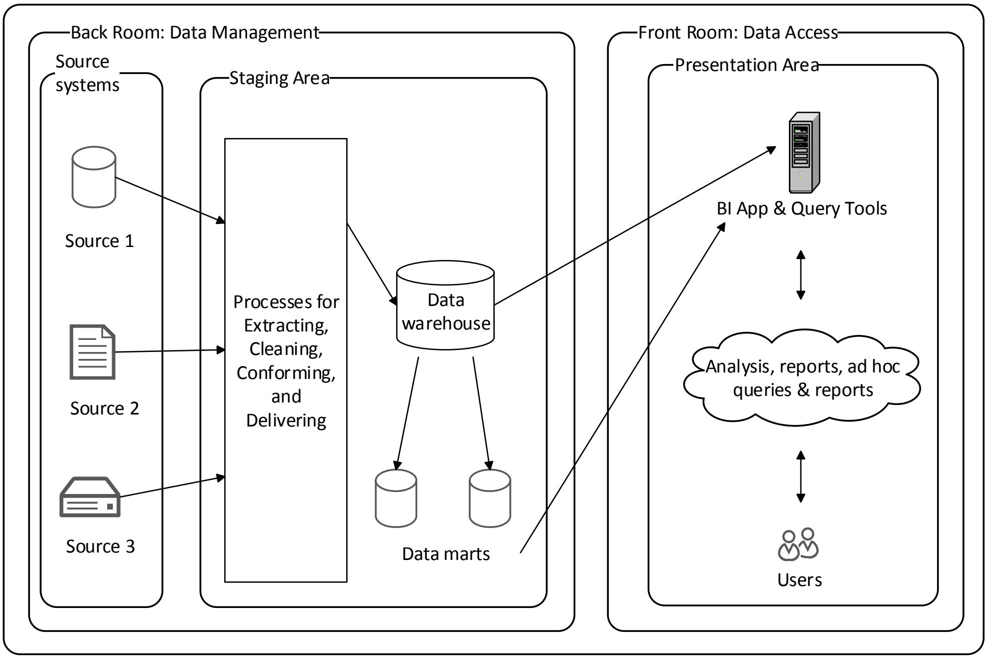


Figure 5. Full data warehouse architecture [12].

The full architecture of the data warehouse is presented in Figure 5. The data warehouse consists of two main parts: a front room and a back room. They are physically, logically, and administratively separate. In other words, the back room and front room are (in most cases) on different machines, they depend on different data structures, and they are managed by different IT personnel. Preparing the data involves acquiring data and transforming it into information, ultimately delivering that information to the query-friendly front room. That approach to data warehousing assumes that data access is prohibited in the back room and, therefore, the front room is dedicated to servicing this single purpose [12]. The back room consists of a set of sources and a staging area. Source systems are the operational systems of record that capture the transactions of the business. The source systems should be thought of as outside the data warehouse. The main priorities of the source systems are processing performance and availability. The staging area consists of a set of processes commonly referred to as extract-transformation-load (ETL), the data warehouse, and (perhaps) several data marts. A data mart is a subset of the data warehouse that is usually oriented to a specific business line or team [13].

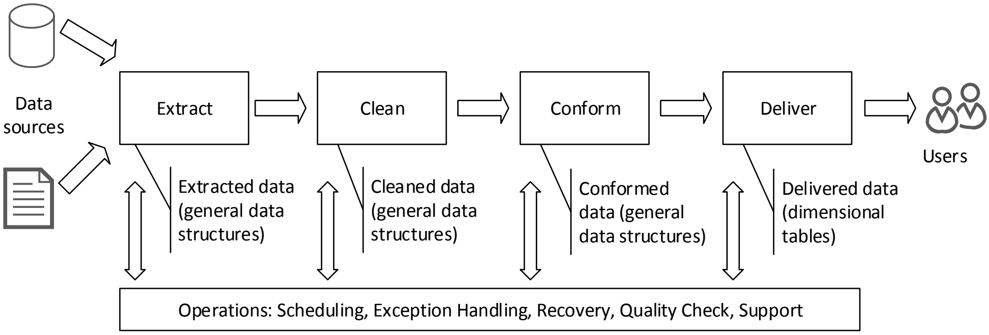

Data sources
Extract
Clean
Conform
Deliver
Users
Extracted data (general data structures)
Cleaned data (general data structures)
Conformed data (general data structures)
Delivered data (dimensional tables)
Operations: Scheduling, Exception Handling, Recovery, Quality Check, Support

Figure 6. The four staging steps of a data warehouse [12].

Data for mentioned storages are staged in four steps, shown in Figure 6. These steps are performed by ETL tools, which implement the ETL process. These four steps are [12]:

- Extracting. The raw data coming from the source systems is usually written directly to the disk with some minimal restructuring. Data from structured source systems often is written to flat files or relational tables. Initially-captured data can then be read multiple times as necessary. In some cases, initially-captured data is discarded after the cleaning step is completed; in other cases, the data is kept as a long-term archival backup.
- Cleaning. In most cases, the level of data quality acceptable for the source systems is different from the quality required by the data warehouse. Data quality processing may involve many discrete steps, including checking for valid values, ensuring consistency across values, removing duplicates, and checking whether complex business rules and procedures have been enforced. Data-cleaning transformations may even involve human intervention and the exercise of judgment. The results of the data-cleaning step are often saved semi-permanently, because the transformations required are difficult and irreversible.

- Conforming. Data conformation is required whenever two or more data sources are merged in the data warehouse. Separate data sources cannot be queried together unless we solve all of the problems of syntactic- and semantic-data heterogeneity.
- Delivering. The whole point of the back room is to make the data ready for querying. The final step is to transform the data structure in order to meet this functionality. This is often a set of simple symmetrical schemes known as multi-dimensional models. These models are a required part of the so-called OLAP-cube construction.

ETL tools play a crucial role in data processing in data warehouses. They provide flexible functionalities of transformation, cleaning, and data quality assurance. These functionalities are designed to provide a useful form of data for end-user applications. ETL technology with its range includes ETL process, methods of design, and modeling the process (as well as the tools that implement it). An example of ETL process is shown in Figure 7.

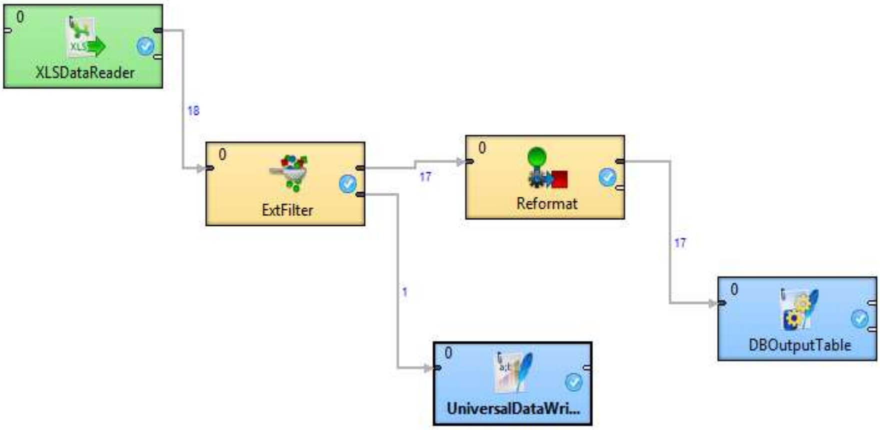


Figure 7. An example of simple ETL process.

Intuitively, such process can be thought of as a directed acyclic graph, with activities and record sets representing the nodes of the graph and input-output relations between nodes representing the edges of the graph [21]. As one can observe, an ETL process is the synthesis of individual tasks that perform extraction, transformation, cleaning, or loading of data in an execution graph - also referred to as a workflow. Also, due to the nature of the design artifact and the user interface of tools, an ETL process is accompanied by a plan that is to be executed [20].

Although, ETL logic is not novel in computer science, several issues still remain open. A main open problem in the so-called traditional ETL is the agreement upon a unified algebra and/or declarative language for the formal description of ETL processes. The optimization of the whole ETL process (but also of any individual transformation operators) pose interesting research problems. In this context, parallel processing of ETL processes is of particular importance. Standardization is a problem that needs an extra note of attention. ETL functionality expands into new areas beyond the traditional data warehouse environment; such cases include (but are not limited to) [21]:

- On-Demand ETL processes that are executed sporadically (typically for web data), and they are manually initiated by some user demand;
- Stream ETL that involves the possible filtering, value conversion, and transformations of incoming streaming information in a relational format;
- (near) Real-Time ETL that captures the need for a data warehouse containing data as fresh as possible.

Finally, with the evolution of the technology and the broader use of the Internet, the interest is moved also to multiple types of data, which do not necessarily follow the traditional relational format. Thus, modern ETL applications should also handle novel kinds of data (XML, websites, spatial, biomedical, or multimedia data) efficiently.

6. Description of proposed architecture

The proposed solution belongs to an update-driven approach. The main idea is to use some elements from architectures of a data warehouse as well as mediated systems. A key role in the proposed solution plays ETL technology, which occurs in two forms: ETL engine and designer. ETL tools provide a flexible design and management of ETL processes, which are a method to solve the data integration problem and allow us to obtain the desired quality of data. However, they have deficiencies in handling any types of data source; for instance, websites. These sources require special functionalities for data extraction. Therefore, the solution has a wrappers layer, which occurs in the mediated systems. Integrated data and meta-data are stored in a relational database. It is also possible to use another kind of database system, like an in-memory database. Queries of users are processed by the business logic layer as mediator. The above-mentioned elements (ETL engine, wrappers layer, global database, and mediator) are the main elements of the proposed architecture, which has been verified through the implementation of a prototype.The solution meets the following assumptions:

- The system provides flexible and independent design and management of ETL processes in order to solve the data integration problem.
- The system allows us to adapt any data source and is extensible in this respect.
- The system provides access to integrated data and methods of creating connections between specific data elements.
- The system automates the creation of connections between data elements.
- The system can be integrated with any presentation layer.

The architecture of the proposed solution is a three-tier architecture, shown in Figure 8. The data layer includes: data sources, source wrapper layer with wrappers, ETL layer with work-flow graphs, and a relational database. Integrated data described by reconciled global schema, meta-data and the relations (connections) between the data elements are stored in a relational database. Data sources are set which is subject-oriented. The wrappers are associated with each selected data source, and their main task is the extraction of appropriate data when the utilized ETL engine can't do that. The results of the wrappers are stored in files, which are the input for the layer based on the ETL technology. In this layer, designed graphs that represent the ETL workflows for each source are executed. Processed data, the connections between them and the meta-data are stored in the corresponding relations of the global schema in the relational database, which is the result of the data integration process.

A global database is available for the business logic layer. This layer contains the implementation of the design pattern called the facade, which is used to standardize access to the system. The facade is, therefore, the access interface for application of the presentation layer. The data layer and the business-logic layer have been implemented as one component (part) that carries out the process of data integration, allows for the implementation of analytical methods, and can be integrated with any presentation layer. This is unlike business intelligence environments, where two separate parts for data integration and analytical methods exist. Therefore, the proposed solution is a data integration system from the perspective of the data layer – but from the general perspective, the proposed solution is a system that implements knowledge discovery from the data process. Integration with any presentation layer means appropriate changes in the interfaces and objects of the business layer. This division provides the ability for independent implementation of the graphical user interface in various technologies. The business logic layer also contains objects responsible for interacting with the global schema by executing the queries. Details of query execution, with regards to application in the foundry industry, will be presented in the next section. Such queries perform some specific functionality described in the facade. Stored thematic aspects, the specific aspects data, and the connections between specific aspects are represented in the logic layer using appropriate objects. This allows the association of the relevant features with these elements. These objects are the data-types system in the proposed solution.

The presentation layer consists of access applications that are accessed by users. It is assumed that access to data and functionalities is achieved by the access applications. As already mentioned, these applications use the facade as an interface.

With respect to users of the system, the following functional requirements were specified:

- browsing integrated data;
- the ability to create connections between specific data elements;
- browsing data connections created by users and the system;
- the ability to delete and update integrated data.

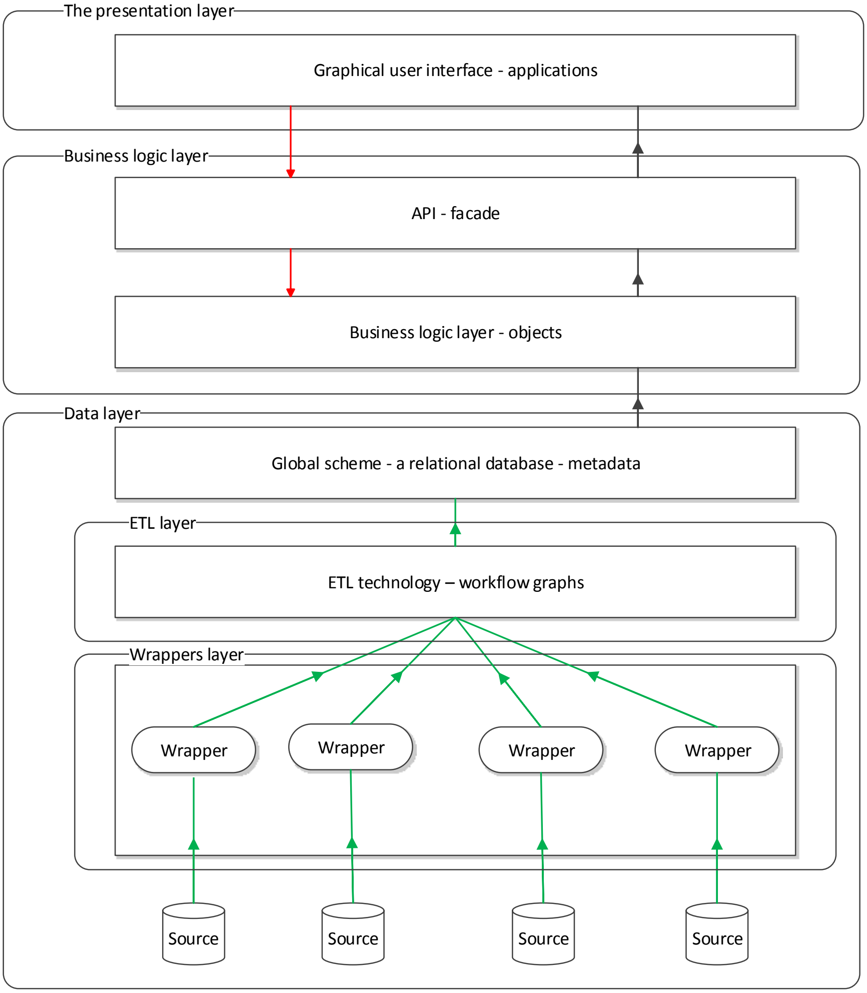

The presentation layer
Graphical user interface - applications
Business logic layer
API - facade
Business logic layer - objects
Data layer
Global scheme - a relational database - metadata
ETL layer
ETL technology – workflow graphs
Wrappers layer
Wrapper
Wrapper
Wrapper
Wrapper
Source
Source
Source
Source

Figure 8. Proposed solution architecture.

In the figure of architecture, user calls to business logic are indicated with red arrows. The black arrows illustrate the flow of data as result of the user call. The format of returned data can be different (for instance, a different level of aggregations) and is determined by the business logic layer. The green arrows illustrate data integration and update flow. The user gains access to the data on demand, and the data is returned from the global schema. The update of such data as well as the first gathering take place at the request of the user. The result is that data stored in the global schema may become out of date. The proposed solution can also perform cyclic data updates.

The wrapper layer includes wrappers which are elements that can be implemented in high-level programming language. They can be separate components of the system and are an implementation of a specified interface. As already mentioned, the wrapper's basic task is to extract and write data to CSV files. Access to these kind of files is supported by most ETL tools and, if necessary, this format can be changed. Wrapper specifies the meta-data, which include:

- with which aspect data source is related;
- name of table in which data will be stored;
- name of ETL workflow graph file associated with a source;
- descriptions of the attributes and their parameters.

The wrapper is the first element that we need to implement in order to attach a data source to the system. However, this requires programming skills. All wrappers as well as the whole extraction process are managed by one object. This process is multi-threaded and, after successful execution, the next phase is performed based on the ETL. If changes are made in data source structure, wrappers have to be adapted to them by changing their source code.

However, wrappers allow us to attach any data source through the use of existing tools or programming libraries and the implementation of additional algorithms. A wrapper can use web-scraping techniques to handle websites as data sources. This gives the designer great flexibility and the ability to obtain and provide already at this level a particular data quality. Wrappers can therefore be programmatically extended by providing such functionalities as detecting changes in the source, searching data sources, etc. The functionality of creating wrappers in a semi-automatic way is also possible.

The ETL layer uses an ETL tool (in the case of the implemented prototype: CloverETL), which provides the ability to embed it in source code and includes an element for visualization, design, and management of ETL processes. It is assumed that these processes can be designed and tested independently and can be delivered to the system as disk file. For each source in the system, an ETL workflow graph is created. Graphs are then executed by the tool's engine embedded in the solution. Graphs determine procedural mapping between specific local schemas and global unified schema. Local schemas are described by meta-data available in the ETL tool. They refer to previously-described CSV files, which are the result of the wrappers layer. Each graph provides the following functionalities:

- extraction, cleaning, transformation, and data loading contained in data integration process;
- possibility of automating the creation of connections between data elements when source describes more than one aspect;
- data update and deleting outdated data;
- identification of data elements through universal and unique identifier inside table.

In contrast to the wrapper layer, we do not need to have programming knowledge to make changes to ETL processes. These tools can be operated via a graphical user interface by the operator of the solution. This implementation provides great flexibility as well as possibilities for development, testing, and management of the ETL process in the solution.

In summary, the main idea is to use some elements from architectures of the data warehouse and mediated systems. A key role in the proposed solution plays ETL technology, and that's why the proposed solution belongs to an update-driven approach. The deficiencies in handling any type of data source complements the wrapper layer, which occurs in the mediated systems. The main task of the solution is to solve data integration problems using the ETL processes. The solution also enhances semantic relations between data elements through the ability to define connections. These links can be created by the user and automatically by the system using ETL. The whole solution consists of two elements. The first one is data integration and mediator component, while the second is a graphical user interface which provides access to functionalities of the first one. It is assumed that the solution can handle any data source.

7. Description of practical application

As mentioned, the data integration system is also an information system which provides useful information from subject-oriented data. The prototype is used to integrate data related to the foundry industry. The system provides information about some aspects of this industry. It is an exemplary application, and the proposed solution can be adjusted to other subject-oriented data.

Mostly, data contained in a particular source refers to a certain aspect of the theme. It is also possible that the data source describes more themes. This gives us the opportunity to obtain a larger amount of data and semantic connections between them. Appropriately-selected data sources and the quality of data included in them are crucial features with regard to benefits that the user can achieve. Specialized data sources are extensive as well as a valuable source of information or knowledge. In the case of the prototype and the foundry industry, two specialized databases from the Foundry Institute in Krakow have been selected. Because of availability and ability to access large amounts of data also Internet sources (websites) have been selected. A full list of data sources that were used to implement the prototype consists of the following items:

- `http://baztech.icm.edu.pl/`. BazTech is a bibliographic online database containing citations from Polish technical journals on engineering, technology, sciences, and the environment. The source contains data that in the system refers to the publications.
- `http://enormy.pl/`. It is an online service that allows an advanced search of standards, not just by title or issue of standards, but also by words that the user

would like to find in the text. The source contains data that, in the system, refers to standards.

- `http://baza-gus.pl/`. It is the complete online database of companies and institutions in Poland. This is the main base of the statistical office, whose task is to collect and share information on most areas of public life. The source contains the data that, in the system, refers to companies.
- `http://www.baza-firm.com.pl/Odlewnie/`. It is a service that contains data related to companies. As in the previous case, data refer to companies.
- `http://www.metale.org/`. This is a metal industry portal which includes information about companies and products. The source refers to two aspects and has been used to obtain information about connections between them.
- SINTE database from Foundry Institute in Krakow. SINTE is a bibliographic foundry database containing abstracts of over 36,000 foundry journal articles (national and international), congresses, and research work. The source contains the data that, in the system, refers to publications.
- NORCAST database from the Foundry Institute in Krakow. NORCAST contains current data on more than 4,000 national and international foundry standards. The source contains the data that, in the system, refers to standards.

Thus, the data from the above-mentioned sources refers to the following aspects of the foundry industry:

- publications,
- standards,
- companies,
- products.

The system takes into account connections between specific data elements created by the user, so it is possible to express relations between standards and a particular company, which can mean that the company meets specified standards. Another example is the link between publication and the product, which may mean that the publication in some way relates to the product. These links provide useful information to users, and their realization can be automated by creating an appropriate ETL process when the data source refers to several aspects and includes appropriate information on a connection between them.

Figure 9 depicts mappings between data sources that relate to certain aspects of the theme as well as tables in which the data will be stored in a global scheme. Aspects can be described by different sets of attributes, depending on the source. For this reason, with every aspect described by the particular source, a table in the global scheme of appropriate nomenclature and structure is associated. Such a realization of storing data related to specific aspects allows easy adjustment of the specified table in the global database if the local source's scheme changes. It also helps to avoid the storage of null values, as it would be in the case when all data related to a particular aspect would be stored in a single table. Results of wrappers and ETL layers, metadata, integrated data, and connections between specific elements will be materialized into global schema. In relation to the user's query referring to some aspect of the theme or connections associated with it (i.e., standards as shown), the business logic layer based on meta-data and links obtained from global scheme will execute queries, combine results, and return them to the user.

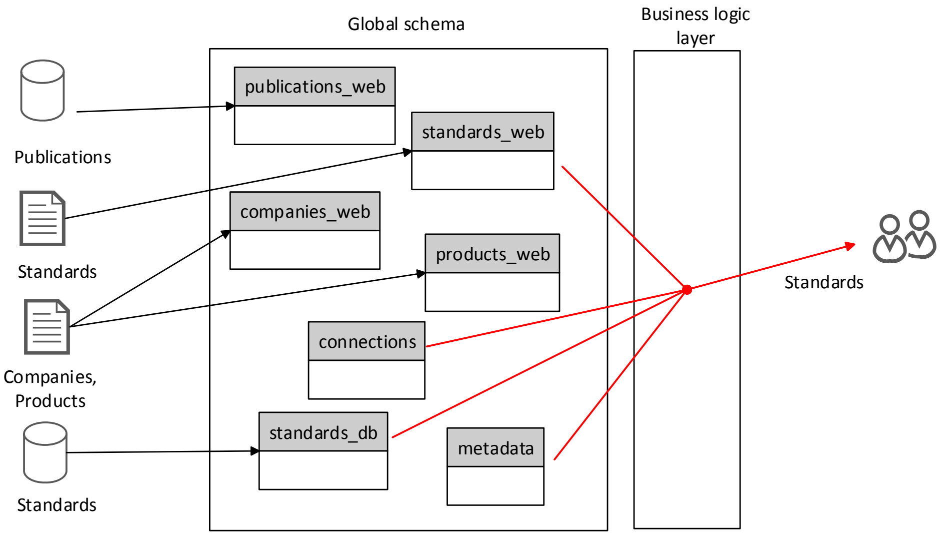


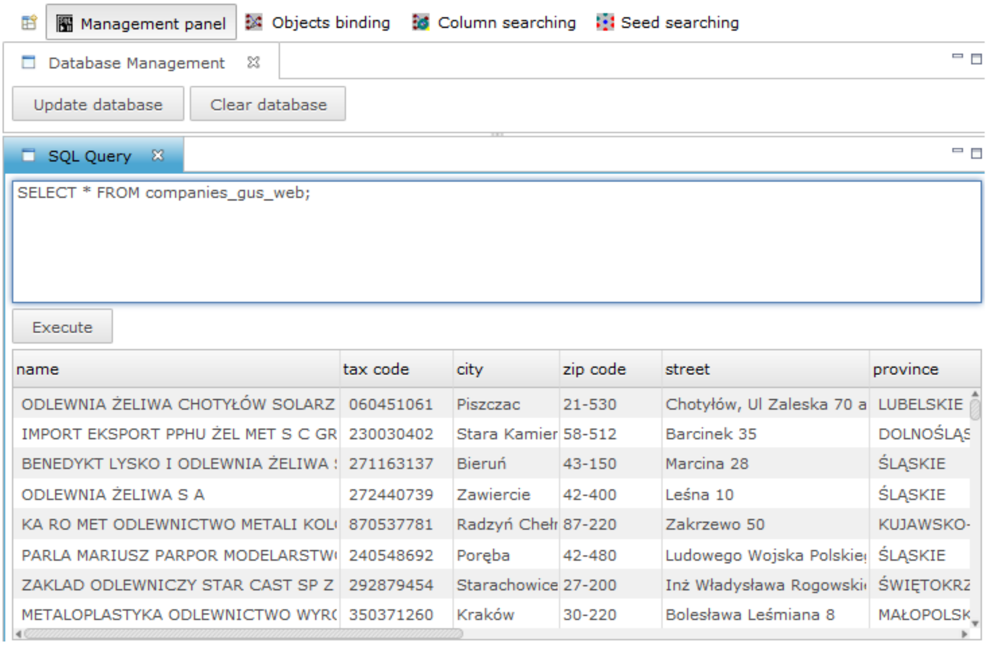
Figure 9. Mappings between thematic aspects and global schema.

8. Verification result

A graphical user interface of the prototype is a RAP web-based application. RAP is a set of tools and widgets that can be integrated with OSGi technology, which is a component environment for the Java language. The prototype as WAR file can be deployed in a JEE web application container (i.e. Apache Tomcat).

The graphical user interface is divided into four perspectives. The first one is a management panel. It provides an update (including integration) and deletion of data and query functionality on a global database. This perspective is shown in Figure 10. After the first start of the prototype, the global database is empty. The user must to press the 'update database' button to start the data integration process and then can start working with data using other perspectives. After initial data load, the user can update data in the global database (with integration) by pressing the 'update database' button. The user can also execute SQL queries on the global database. A sample query shown in Figure 10 returns all attributes of companies_gus_web table, which refers to companies.

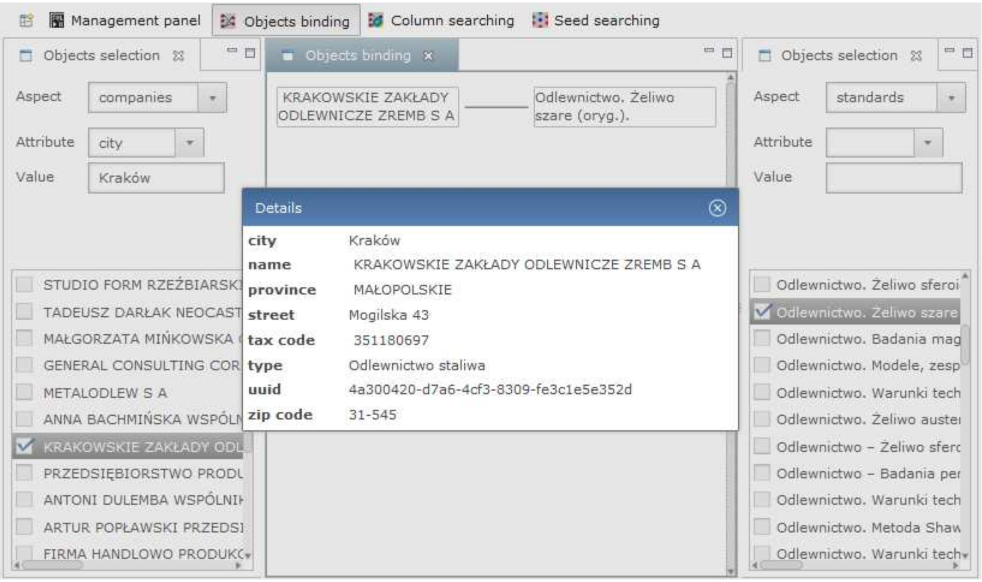
The next perspective is a perspective of object binding. It is shown in Figure 11. It allows us to browse integrated data associated with certain thematic aspects and to create connections between them.



```
SELECT * FROM companies_gus_web;
```

name	tax code	city	zip code	street	province
ODLEWNIA ŻELIWA CHOTYŁÓW SOLARZ	060451061	Piszczac	21-530	Chotyłów, Ul Zaleska 70 a	LUBELSKIE
IMPORT EKSPORT PPHU ŻEL MET S C GR	230030402	Stara Kamier	58-512	Barcinek 35	DOLNOŚLĄS
BENEDYKT LYSKO I ODLEWNIA ŻELIWA :	271163137	Bieruń	43-150	Marcina 28	ŚLĄSKIE
ODLEWNIA ŻELIWA S A	272440739	Zawiercie	42-400	Leśna 10	ŚLĄSKIE
KA RO MET ODLEWNICTWO METALI KOL	870537781	Radzyń Chełr	87-220	Zakrzewo 50	KUJAWSKO-
PARLA MARIUSZ PARPOR MODELARSTW	240548692	Poręba	42-480	Ludowego Wojska Polskie	ŚLĄSKIE
ZAKLAD ODLEWNICZY STAR CAST SP Z	292879454	Starachowice	27-200	Inż Władysława Rogowski	ŚWIĘTOKRZ
METALOPLASTYKA ODLEWNICTWO WYR(	350371260	Kraków	30-220	Bolesława Leśmiana 8	MAŁOPOLSK

Figure 10. Management panel perspective.



Aspect: companies
Attribute: city
Value: Kraków

KRAKOWSKIE ZAKŁADY ODLEWNICZE ZREMB S A — Odlewnictwo. Żeliwo szare (oryg.).

Aspect: standards
Attribute:
Value:

city	Kraków
name	KRAKOWSKIE ZAKŁADY ODLEWNICZE ZREMB S A
province	MAŁOPOLSKIE
street	Mogilska 43
tax code	351180697
type	Odlewnictwo staliwa
uuid	4a300420-d7a6-4cf3-8309-fe3c1e5e352d
zip code	31-545

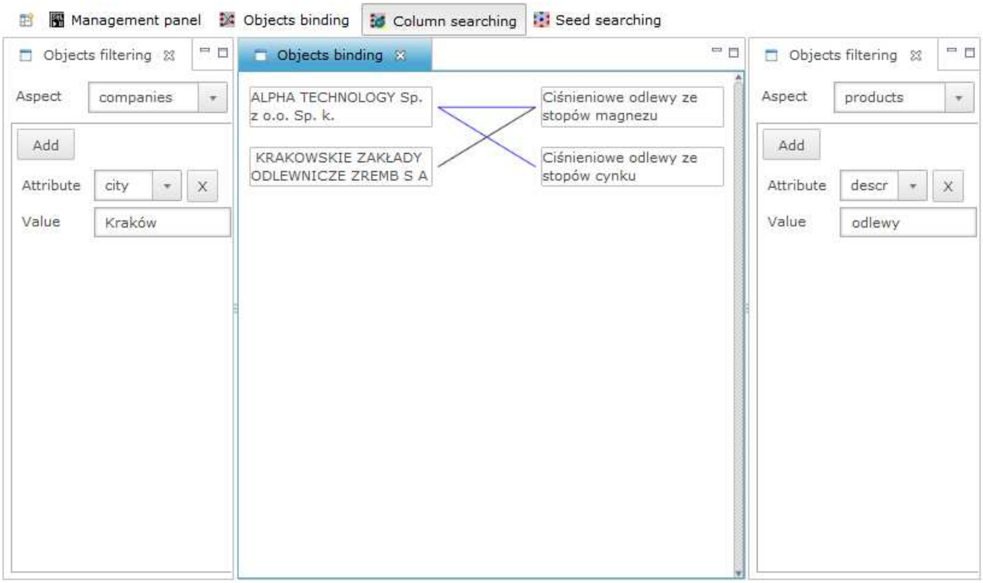
- □ STUDIO FORM RZEŹBIARSKI
- □ TADEUSZ DARŁAK NEOCAST
- □ MAŁGORZATA MIŃKOWSKA (
- □ GENERAL CONSULTING COR
- □ METALODLEW S A
- □ ANNA BACHMIŃSKA WSPÓLN
- ☑ KRAKOWSKIE ZAKŁADY ODL
- □ PRZEDSIĘBIORSTWO PRODU
- □ ANTONI DULEMBA WSPÓLNIK
- □ ARTUR POPŁAWSKI PRZEDSI
- □ FIRMA HANDLOWO PRODUKC

- □ Odlewnictwo. Żeliwo sferoi
- ☑ Odlewnictwo. Żeliwo szare
- □ Odlewnictwo. Badania mag
- □ Odlewnictwo. Modele, zesp
- □ Odlewnictwo. Warunki tech
- □ Odlewnictwo. Żeliwo auster
- □ Odlewnictwo – Żeliwo sferc
- □ Odlewnictwo – Badania per
- □ Odlewnictwo. Warunki tech
- □ Odlewnictwo. Metoda Shaw
- □ Odlewnictwo. Warunki tech

Figure 11. Objects binding perspective.

Data can be searched using filters related to the given aspect's attribute. The figure depicts integrated data of companies located in Krakow and data relating to standards. After selecting the desired company, it shows up in the central part and provides a view of detail data. The same applies to the choice of the wanted standard. After selecting these two aspects, it is possible to realize the connection there between. It is represented by a line. These connections represent semantic relations of data and are valuable information for system users.

The next perspective is a perspective of column search. It is shown in Figure 12. It is used to search for connections between data elements. The figure depicts a situation in which the user is searching for companies located in Krakow and which have a castings in offer. Three connections have been defined. They are divided into two types marked with different colors. Blue represents connections automatically created in the ETL phase of data integration, while black represents connections created by the user as a result of work with the prototype. Two companies that have in their offer castings have been found.



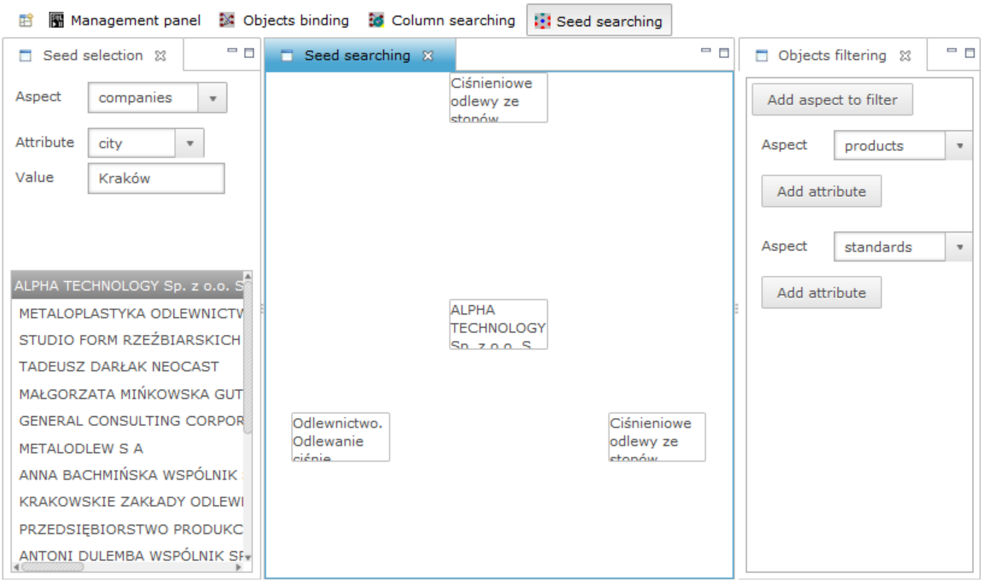
Aspect companies
Attribute city
Value Kraków

ALPHA TECHNOLOGY Sp. z o.o. Sp. k.
KRAKOWSKIE ZAKŁADY ODLEWNICZE ZREMB S A
Ciśnieniowe odlewy ze stopów magnezu
Ciśnieniowe odlewy ze stopów cynku

Aspect products
Attribute descr
Value odlewy

Figure 12. Column search perspective.

The last perspective is a search from the grain perspective. It is shown in Figure 13. It allows us to search connections between a particular aspect and greater number of other aspects. The figure depicts an example of a link search between companies from Krakow and two other aspects, which are products and standards. Of course, the user can filter the aspect's attributes specifying a desired connection. The selected company will be represented in the central region of perspective, and related objects will be around it. The user can browse details of any visible data object. The selected company has a connection with two products, which are castings, and one connection with standards. These connections may mean that the company has in its offer two types of castings and meets standard for pressure casting.



Aspect companies
Attribute city
Value Kraków

ALPHA TECHNOLOGY Sp. z o.o. S
METALOPLASTYKA ODLEWNICTV
STUDIO FORM RZEŹBIARSKICH
TADEUSZ DARŁAK NEOCAST
MAŁGORZATA MIŃKOWSKA GUT
GENERAL CONSULTING CORPOR
METALODLEW S A
ANNA BACHMIŃSKA WSPÓLNIK
KRAKOWSKIE ZAKŁADY ODLEWI
PRZEDSIĘBIORSTWO PRODUKC
ANTONI DULEMBA WSPÓLNIK SF

Aspect products
Aspect standards

Figure 13. Searching from the grain perspective.

As shown, the implemented prototype is an information system relating to the foundry industry. It provides various functionalities for users to obtain useful information. Then, interpretation of such information can provide knowledge.

9. Conclusion

The implemented prototype is the proposed solution's proof of concept. The proposed architecture solves the data integration problem, which was the main goal. It performs syntactic integration with ETL workflows, and supports semantic integration through connections between data elements. Creating these links can also be automated using ETL workflows. Usage of the ETL tool enables flexible management of ETL processes responsible for data integration. It is also possible to add any data source through the implementation of appropriate wrappers. The prototype is therefore extensible, but this process is complicated and requires programming skills. The same applies in adaptation to changes in the local schemes due to the autonomy of data sources. Further development may involve the expansion of layers in the proposed architecture; for example, expansion in the direction of a data virtualization server, semi-automatic wrappers creation, or appropriate data source searching.

References

[1] Calvanese D., Giacomo G.D., Lenzerini M., Nardi D., Rosati R.: Source Integration in Data Warehousing. *DEXA Workshop*, 1998.

[2] Calvanese D., Giacomo G.D., Lenzerini M., Nardi D., Rosati R.: A Principled Approach to Data Integration and Reconciliation in Data Warehousing. In: *Proceedings of the International Workshop on Design and Management of Data Warehouses*, 1999.

[3] Calvanese D., Giacomo G. D., Lenzerini M., Nardi D., Rosati R.: Data Integration in Data Warehousing. *Int. J. Cooperative Inf. Syst.*, 2001.

[4] Doan A., Halevy A., Ives Z.: *Principles of Data Integration.* Morgan Kaufmann, 2012.

[5] Halevy A. Y., Rajaraman A., Ordille J. J.: Data Integration: The Teenage Years. *VLDB*, 2006.

[6] Han J., Kamber M.: *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2012.

[7] Hull R., Zhou G.: *A Framework for Supporting Data Integration Using the Materialized and Virtual Approaches.* 1996.

[8] Inmon W. H.: *Building the Data Warehouse.* Wiley Publishing, Inc., 2005.

[9] Ives Z. G.: Efficient Query Processing for Data Integration. *A dissertation for the degree of Doctor of Philosophy*, 2002.

[10] Kermanshahani S.: Semi-materialized framework: a hybrid approach to data integration. *ACM*, 2008.

[11] Kermanshahani S.: *IXIA (IndeX-based Integration Approach) A Hybrid Approach to Data Integration.* A dissertation for the degree of Doctor of Philosophy, 2009.

[12] Kimball R., Caserta J.: *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Wiley Publishing, Inc., 2004.

[13] Kimball R., Ross M.: *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* Wiley Publishing, Inc., 2002.

[14] Koch C.: *Data Integration against Multiple Evolving Autonomous Schemata.* CERN-THESIS-2001-036, 2001.

[15] Kurgan L. A., Musilek P.: A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, 2006.

[16] Lenzerini M.: Data Integration: A Theoretical Perspective. *PODS '02 Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002.

[17] Levy A. Y.: The Information Manifold Approach to Data Integration. *IEEE Intelligent Systems*, 1998.

[18] Negash S.: Business Intelligence. *AMCIS*, 2003.

[19] Tatbul N., Karpenko O., Convey C., Yan J.: Data Integration Services. *Brown University, Computer Science*, 2001.

[20] Vassiliadis P.: A Survey of Extract-Transform-Load Technology. *Integrations of Data Warehousing, Data Mining and Database Technologies*, 2011.

[21] Vassiliadis P., Simitsis A.: *Extraction, Transformation, and Loading. Encyclopedia of Database Systems*, 2009.

[22] Vercellis C.: *Business Intelligence: Data Mining and Optimization for Decision Making.* A John Wiley and Sons, Ltd., 2009.

[23] Widom J.: Research Problems in Data Warehousing. In: *Proceedings of International Conference on Information and Knowledge Management*, 1995.

[24] Wiederhold G.: Mediators in the architecture of future information systems. *IEEE COMPUTER*, 1992.

Affiliations

Marek Macura
AGH University of Science and Technology, Krakow, Poland

Received: 14.06.2013
Revised: 25.11.2013
Accepted: 3.12.2013