EXTRACTION OF INFORMATION IN THE CONTEXT OF BUSINESS INTELIGENCE

Branko Latinović

Pan-European University APEIRON, Banja Luka

General survay

DOI: **10.7251/JIT1101054L** UDC **005.94:004.738**

Summary: business Intelligence in the developed business systems allows better reasoning and decision making. ETL processes represent the most important processes in the system of Business Intelligence. It is about extracting, transforming, and filling a Data Warehouse with data which then transforms into data that is by its nature new and presented in a way that is meaningful and useful in an actual business organization. In conjunction with the methods of Information Extraction, knowledge is significantly expanded and given a completely new image. Intention is the collection of data that is available and processing the same in one place, regardless of whether the data was in a structural form or any other.

Key words: Business Intelligence, Data Warehouse, Information Extraction

INTRODUCTION

Information technologies today, represent a dominant infrastructural way in almost all spheres of social life. With that, collection, safe-keeping, processing and use of data and information, give modern management a whole new character. These activities are creating functional knowledge for the decision making process.

The business is based on data which is transformed into information and knowledge. Business organizations transform information into knowledge, into business solutions in several ways. The process starts with collecting data from various sources and storing it into database, then selection and processing of that data in order to be in a format that fits the data warehouse. Then users are using data from a data warehouse for analysis. The analysis is done by using the analysis tool that searches for patterns and by using intelligent systems that support the interpretation of data [6].

From the existing data sources which can be transactional data sources, OLAP cubes, various ERP and CRM solutions, as well as texts which are located on local PCs , data is extracted, transformed and filled with so called ETL processes into analytically oriented systems or data warehouses. Data warehouse is a system which is with its structure modified to fit analysis and business concept of its users, and it is not that dependent on the platform and base type, in which it will be implemented. After the data base construction, the report system and analytics are built.

Contribution of the information extraction methods to the business intelligence is clear in itself. It is about new information that with these remakes is becoming available and which further contributes to the development of final reports and the eventual making of the decision that should be the result of the whole data processing whether it is the case of structured data or unstructured data. Less clear is the impact that methods of deep data analysis can have on the systems for information extraction. Finding different non-obvious connections between data acquired from textual documents, one can come up with new findings and ideas in regards to what type of information is even useful to search for in the text, and with that, a sum of extracted information which is useful is growing over time.

BUSINESS INTELLIGENCE

Business Intelligence (short BI) is a set of methods and software tools which enables use of data from the data warehouse (short DW) and its transformation into information needed for the business decision making.

Business intelligence system is such system which saves information and knowledge about competition, buyers, suppliers and processes. It allows business negotiations and reasoned presentation towards buyers and suppliers, quality operational planning, competition behavior tracking, certain market segments' overview, and future events forecast. Besides stated, business intelligence system offers better insight into understanding of existing buyers and knowledge into what stimulates them to behave in a certain way.

Business intelligence started developing intensively when business organizations automated their business processes, i.e. when they implemented different transactional systems, which have very soon proved as generators of large amount of information. From the technical point of view, business intelligence is a process with which raw data is transformed into information. Such information is then analyzed and used in the decision making process within the organization.

Conducting business intelligently means introduction of a business intelligence concept deep into the existing organization's structure. This raises a question: In which way can the business intelligence be successfully integrated into business processes of the organization so that employees could at any point in time use it and give their personal input in the realization of the strategic business goals. Every next point, every next business intelligence implementation project will rely on strategic goals which have been identified in this initial step, and which are documented in the business strategy of the organization.

When organization's strategic goals are in question, it is necessary to have an absolute agreement on their definition, and each one individually, represents the base for a potential business intelligence project. Employees must come to an agreement in regards to priorities related to the strategic goals, so that it is clearly defined which one will be chosen as a base for the initiative in the field of business intelligence.

One of the strategic goals can, for example, be "operational costs reduction". On the basis of this example question can be raised: "how does business intelligence help in the operational costs reduction?"

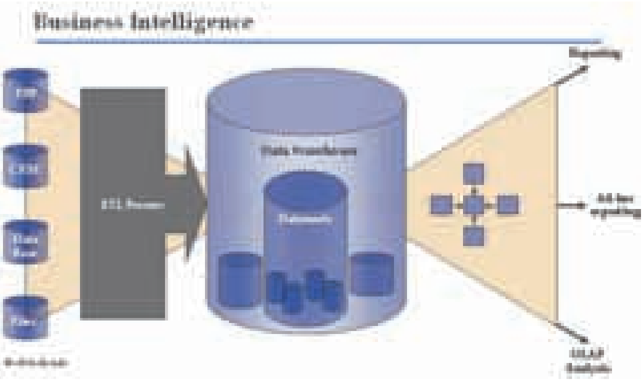
Key is in the identification of the business intelligence service in every process in order to achieve set goals, then in the business intelligence integration into these processes, and at the end, it must be taken into consideration that one business process can be intertwined through more organizational parts of the organization.

Because of that, it is not only important to understand the business process, but also the role of the people involved in that process, as well as applications which they use in each of their activities. Such approach is necessary in order to understand the mechanism of business intelligence integration into business processes.

Some activities in business processes happen automatically and are controlled by software, while others are done manually and are controlled by people, participants in the process [1]. In automated activities business intelligence is integrated directly through some technology, i.e. web service technology. In case the activity is controlled by an individual, some issues must be addressed. Primary issue is the role that each individual has, as well as the application that individual in question is using in the development of a certain activity of the business process.

Also, there are some technical questions as well, for example, is the individual in a remote location and will it access business intelligence through a mobile device? Role identification is equally important so that proper business intelligence form is applied in the context of specific business activity which participant with certain role, is conducting, and at specified time. Business intelligence system architecture, Picture 1.

PICTURE 1. BUSINESS INTELLIGENCE SYSTEM ARCHITECTURE



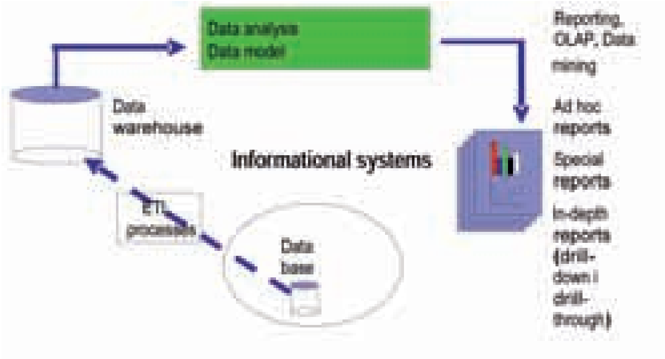
DATA WAREHOUSE

Today, every organization or a company disposes not only with a large amount of information, but also with a large number of data sources, hence the need for the integration of all possible systems. Big step in this regard has been made with the introduction of data warehouse which should put together all data within the company, independent of the type of data or an application, but here all the textual documents are neglected. In order not to, again, come to the separation of the systems due to different types of inputs, there is a need for the integration of information which is available in non-structural or semi-structural text. The goal is to later equally use that information with all other data available in common forms and ways.

As getting the quality information on time is important for gaining the advantage over the competition, the manager must get the same information as fast as possible and in the form which is adjusted to his needs. From that is concluded that from today's companies' informational systems is expected to secure information which content, performance speed and the way it is presented, conforms to immediate needs of the manager in the decision making process. While for the purpose of operational business management, classic data bases are used, based on relational model, and fast, the actual state of the system with certain data, after it is updated, is lost, and for the purpose of making correct business decisions, it is necessary to have insight into business events time line, so such data bases do not represent satisfactory solution.

Due to the above stated phenomena, new ways of information organization in informational system of computer memories are created. Developed is the new generation of computer systems based on the concept of data warehouses. Data warehouse contains information collected from different sources, company's business historical data, as well as data outside of the company, and it is designed in a way that allows data search, on-line analytical processing, reporting and support to the decision making process. Date warehousing process illustrated in Picture 2.

PICTURE 2 DATA WAREHOUSING



Data warehouse is by technical specifications and by content, different from transactional systems based on transaction manipulation. Even though operational base is its assumption, data warehouse, in its design, relies on multidimensional concept. So, the new generation of computer systems now consists of two parts, operational (transactional) and data warehouse (analytical), which achieves separation of the information generating processes (extraction, aggregation, reporting, analysis) which by their nature differ from operational processes.

According to the definition created by Inmon [9], data warehouse represents subject oriented, integrated, time-variant, and non-volatile sum set of data, and it ultimate goal is to help management in the decision making process.

Subject orientation of data means that data is organized around subject, in a way that it gives information about clearly defined subjects within the functional field (i.e. sales, procurement, etc.) instead of about current operations of the company. Opposite to this, operational data bases are organized around business applications, so they are directed towards the current operations (i.e. order processing, deliveries and similar).

Integration means that data is collected into data base, from different sources and it is always stored in the same format, so that it is shown in a consistent manner.

Connection to the time means that all the data in data warehouse is identified in relation to a certain time period, meaning that it has a historical character. As oppose to that, in operational data bases, stored are only current, and the most recent data. However, from the point of view of business intelligence concept, all-inclusive future events forecast is not possible without the history knowledge of the same or some other events. This means that even though data in the data warehouse reflects past, its orientation is towards the future.

Non-volatility requires form of data that it is stable and once stored in the warehouse, by rule, is unchanged. This allows management or anyone who uses data warehouse to be sure to get the same answer regardless of time or frequency.

Data warehousing process represents continuous planning process, data collection from different sources, data use, maintenance management and continuous upgrade. Among many steps in this complex continual process, it is important to emphasize the importance of having a vision of what is to be achieved by creation of a data warehouse. One of the roles of warehouses is development and use of data-based knowledge. Primary function of data warehouses is collection of data and creation of logically integrated and subject-oriented information. Warehouses should be modeled in a way that they could easily and quickly be modified to all the changes and requirements in the business environment.

Taking into consideration subject-orientation of the data, when modeling warehouses applied are techniques which support subject orientation and secure enough adaptability in order to, over time, be able to integrate data from additional sources. Data warehouse should be a source of stable data, independent of eventual changes in the business processes. Free from operational processing, data warehouse secures information generation process upgrade, and through techniques of knowledge discovery secures continuous findings of new information.

ETL process is a term for a data transfer process from data transactional systems into data warehouses and unavoidable link to a development of business intelligence system. The name comes from an English words extract, transform and load. ETL process includes:

- (Extract) collection of data from the outside sources,
- (Transform) data adjustment in line with business needs,
- (Load) data upload into data warehouse.

ETL process is very important because it defines the way of data upload into data warehouse. Term ETL can be used for naming the upload process into any data base.

Collection (Extract)

First phase of the ETL process is data collection from different system sources. Each individual system can be using different organization or data format. There are many different data source formats, and most often are used relational data bases or unrelated data bases. After completing data collection, they are in columns which are sometimes called fields. After this, each type of data can be individually processed.

Transformation (Transform)

Transformation phase refers to a series of commands or functions within collected data which secure data upload. Some data sources do not require complex processing of e-data. In some cases any of the below listed processing combinations may be demanded:

- Upload of only specifically chosen data column
- Recalculation of coded value (i.e. system source stores M as a mark for male and F as a mark for a female)
- Securing the new recalculated value
- Pulling the data from multiple sources at the same time
- Sum of more data rows (i.e. joint sales of all regions)

Upload (Load)

In the upload phase data is uploaded into data warehouse. The scope of the process depends on the size of the company or organization. Some data warehouses exchange old data with new, and complex systems may even store data from the past and track its changes.

Information Extraction

Basically, there are two types of extraction of information, depending on the type of texts that are manipulated. The first relates to the extraction of knowledge and it is possible if the documents themselves contain that knowledge, not just a group of data, which should undergo further processing. The main problem in this approach to information extraction is that the extraction of knowledge is extremely complex because of the language features and demanding methods of processing of the natural language. In addition, most methods can find connections between data that is physically located near one another, but some further cause and effect relationships are much harder to detect. When the extracted information is actually a specific value, we cannot treat it as knowledge, and it is necessary to make another step. This applies to in-depth analysis of extracted data. This process of extraction of information actually serves to convert the text into a structured record of the same information, and shall continue to apply some of the methods of in-depth data analysis or this information is further only treated for the purposes of reporting within the business intelligence [8].

Information extraction types

Extraction types are not strictly determined when it comes to the rules of information extraction from the texts, by information extraction methods. However, there are common types of information extraction which are used in most systems. Those types are based on distinguishing the names, finding of phrases with the same meaning, semantic roles, connections between entities and time periods. All other information is dependent on the specific system and actual textual source of the information [4].

Distinguishing names refers to distinguishing and classification of terms in the text which define some name of the person, organization, place, position, etc. This is the simplest, but also the most reliable information extraction type. When we talk about phrases with the same meaning, then we have synonyms a.k.a. different names for the same person or a thing. In use are also different linguistic elements which reference direct description of an actual object somewhere earlier or later in the text. Good example of such linguistic element are pronouns *he* or *she* which refer to the already mentioned person.

Semantic roles are assigned to different syntactic parts of sentences. They determine some actions or states of participants, and consequences, and can be more or less generalized.

There is also a possibility of detecting the connection or relation between the entities found in previously mentioned methods. Typical examples of such relations would be that such specific person (first entity) works for the specific organization (another entity) and lives in a particular place (the third entity).

Recognizing the time elements takes place in two phases. In the first phase, expressions of time are to be found such as absolute time or a specific date or time, the relative temporal expressions such as, for example, yesterday or tomorrow, the relative terms related to a specific event, periodically repeated, etc. After finding such expressions the timeline can be determined in which certain events occurred.

It should be emphasized that all the information that can be found in the text is independent of the domain of the text, but also could very well describe some event or object. Some events are often also interdependent, so the next step is to connect a series of events into a more complex scenario [5].

Methods

Theory and practice distinguish two basic ways of extracting data such as [3]:

- Knowledge engineering approach or symbolical techniques and
- Automatic trainable systems

With first method it is necessary to have rules, which are mostly created by linguists in cooperation with experts in the field under which system operates. Here, most of the time is spent studying a set of documents and generating and optimizing a set of rules. The problem is that information can be found in different shapes and contexts, which is very hard to predict in advance and take into consideration all the cases with the help of rules. With automatic learning everything is based on statistics and additional linguistic knowledge is not necessary, which means that they are independent of the language of the input documents. The biggest problem with this method is a need for a large set of input documents, which are used for studying in order to achieve rationality of the system. These methods are much more efficient and there is a significant number of learning methods which can be applied to this type of information extraction [7].

Symbolical techniques

During the knowledge engineering which is still used for information extraction, often are used regular grammar and regular expressions, which do not permit the emergence of elements that are not final. It is clear that this is a partial parsing, where taken are only predetermined parts of the text while others are ignored. Regular grammar can very well show patterns in the text, and when implemented using final automat it is very efficient in parsing text. Final automats are often used during the extraction of information, because many samples in the text have fixed, pre-defined order. For a complete natural language processing final automat is not good enough. For this reason we moved from the merger of several final automats in the network, which can be organized in several ways. In any case, final automats have many limitations, but their main advantage is that they are fairly easy to implement and maintain, as a result of a relatively small set of regular expressions with which they are defined. Moreover, the additional advantage of this approach is the high-speed parsing of the text [2].

Automatic learning

Information extraction is often done with the help of automatic learning methods. Since it is the case of detection and recognition of certain information within large amounts of text, it all actually comes down to recognition of samples, i.e. their classification. Samples are recognized based on a combination of features and their values. In this case those features are text characteristics which can be identified and measured. Automatic learning methods replaced manual creation of rules and knowledge creation based on which, information extraction would be performed. Besides, additional advantage is that those methods are not of deterministic nature, but certain element can be added to a certain class with specific probability. Human being actually functions in the same way, because while reading the text we come to many conclusions based on some uncertain information. That uncertainty is also, later on, during further processing of received information, good to take into consideration. [6, 2]

Conclusion

In order for business intelligence systems to really include all available information and the conclusions that can be drawn from them, it is necessary to introduce additional sources of data. One large and poorly explored set is text data, which capabilities with appropriate treatment are extremely high. In this direction one can start moving by using the method for information extraction, where information is what needs to be found and represents a new assignment for which purpose it is necessary to consolidate knowledge in the field of business intelligence, and the processing of text data. Combining these two methods, business intelligence gets a new dimension.

Literature:

[1] Ciric, B. (2006). “Poslovna inteligencija“, Data status, Beograd, pp. 33.

[2] Cunningham, H. (2004). “*Information Extraction, Automatic*”, Elscevier Science.

[3] Feldman, R. (2003). “*Information Extraction Theory and Practice Tutorial*”, Third IEEE International Coonference on Data Mining.

[4] Moens, M. F. (2006). “*Information Extraction: Algorithms and Prospects in a Retrieval Context*”, Springer.

[5] Pazienza, M. T. (1997). “*Information Extraction, A Multidisciplinary Approach to an Emerging Information Technology*”, Springer.

[6] Rainer, R. and Turban, E. (2006). “*Uvod u informacione sisteme*“, Data status, Beograd, pp. 106.

[7] Srbljic, S. (2002). “*Jezicni procesori 1, Uvod u teoriju formalnih jezika, automata i gramatika*”, Zagreb.

[8] Taylor, S. M. (2003). “*Improving Analysis with Information Extraction Technology*”, 8th ICCRTS.

[9] The Business Intelligence and Data Warehousing Glossary”, http://www.sdgcomputing.com/glossary.htm, 2. 4. 2004.

Submitted: April 27, 2011
Accepted: June 02, 2011