Business Intelligence and its Applications to the Public Administration

Dr. Davide Barbieri, University of Ferrara, Italy
Email: davide.barbieri@unife.it

Abstract

Data which have been filtered and analyzed, in order to produce a synthetic, structured and meaningful set of information - made available for strategic decision making - are put under the broad definition of intelligence.

Business Intelligence (BI) consists in gathering intelligence from corporate data. It is a complex process, including many components and functions. The basic ones are: On-line Transaction Processing (OLTP), data warehousing, On Line Analytical Processing (OLAP) and data mining.

OLTP is the process of extracting data from a transactional database. OLTP results can be sets of records or statistical indicators like means and standard deviations. A data warehouse is a multi-dimensional data storage facility. Data warehouses allow performing OLAP: fast analysis of shared multidimensional information. Data mining is a powerful technique for knowledge discovery, based on statistical algorithms. Its aim is to find hidden patterns and models (associations, correlations etc.), together with counter-intuitive information.

The Italian administration has recently adopted a Business Intelligence system to measure its efficiency. The system relies on the measurability of results, comparing them to the expected output. Lower-than-expected efficiency may result in increased auditing and eventual penalization. Good or excellent performances may instead be rewarded.

Different projects have been undertaken in order to prevent or detect frauds, especially in the fiscal and financial areas. Predictive modelling may improve auditing efficiency, reducing the costs related to useless audits and increasing the probability of effective ones.

Keywords: *Business Intelligence, information overload, data mining, efficiency, fraud detection*

Introduction

Humans involved in any kind of competitive activity - be it hunting, trade or war - have always managed to search for information about their prey, rival or enemy. Data concerning eating habits, movements, behavioral patterns, strengths and weaknesses, have been collected and analyzed accurately, then reported to whom they might concern, like team or military leaders and co-workers.

Nowadays though, the quest for knowledge may be overwhelming, mainly because of two reasons: the diversity of sources (television, newspapers, computers, radio, phone calls etc.) and the enormous amount of data, especially those stored in huge corporate databases, traveling through the Internet or through local intranets, allowing for fast communications and frequent updates. Even though state-of-the-art technology provides many advantages in terms of performances, the amount of data available has led to the problem of information overload.

Still, being able to retrieve the proper information at the right time can give any organization a tremendous competitive advantage, especially in the case of a large company or institution, like a ministry or an armed force. Knowledge which has been filtered and analyzed, in order to produce a synthetic, structured and meaningful set of information - made available for strategic decision making – is put under the broad definition of intelligence.

In this paper, we shall see how intelligence is gathered and applied to the business world and the public administration, especially in order to extract from corporate databases the necessary information for decision making.

Business Intelligence as a step-by-step process

Business Intelligence (BI) is the process of retrieving, extracting, filtering and analyzing corporate data in order to produce concise and meaningful information mainly for decision support. This kind of intelligence is usually presented in the form of a written report, summary or presentation, with charts.

The first part of the process consists in gathering information that is structured data. Data are structured when they are saved inside a real database management system. The basic structure usually takes the form of a table, like in relational databases (Microsoft Access and SQL Server are two popular examples) and data must have a defined data type, be it text, number, currency, date etc. Information is then filtered, processed and analyzed in order to produce intelligence (as schematically represented in figure 1).

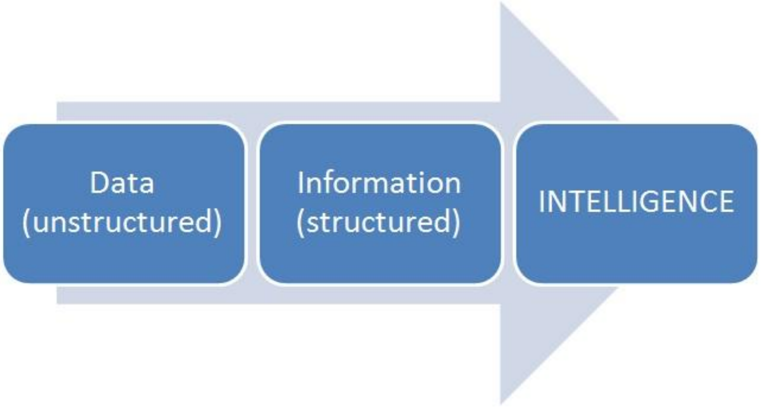


Figure 1: From data to intelligence.

Unfortunately though, most of the times data are unstructured. In fact, companies store most of their data inside Excel spreadsheets, emails, Word documents etc. Therefore, before beginning the actual knowledge discovery process, BI consultants and business analysts (or domain experts), team up with Information Technology staff, in order to assure a comprehensive data structure, including all the data which are of any interest for the organization. Their main task will be to “normalize” the data, reducing redundancy, eliminating inconsistencies and providing unique identifiers for all the entities represented inside the database.

Subsequently, BI consultants usually implement the following functions (the whole process is represented in figure 2):

a. On-line Transaction Processing (OLTP).
b. Extract, Transform and Load (ETL).
c. Data warehousing.
d. On-Line analytical Processing (OLAP).
e. Data mining.

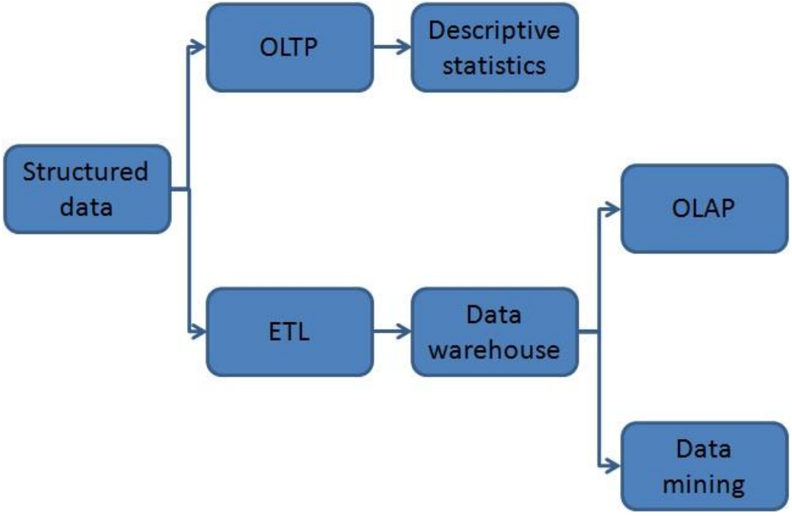


Figure 2: Business Intelligence as a set of functions

On-line Transaction Processing

On-line Transaction Processing is the process of extracting information from a transactional database. It can be performed by means of a query language, like SQL (Structured Query Language). SQL may filter the records inside the database, imposing conditions that must be met in the results.

BI is not extracting data about a single individual, but rather synthetic results about the whole domain of interest. One way to produce meaningful and concise information consists in using a small set of statistical indicators of central tendency and dispersion. Typical measures of central tendency are mean and median, while standard deviation and range can be used as measures of dispersion.

In the case of a big organization, where the databases may contain records about thousands (or millions) of customers, such results could not be reasonably achieved by means of a manual task, but only thanks to automated statistical data analysis, as in figure 3.

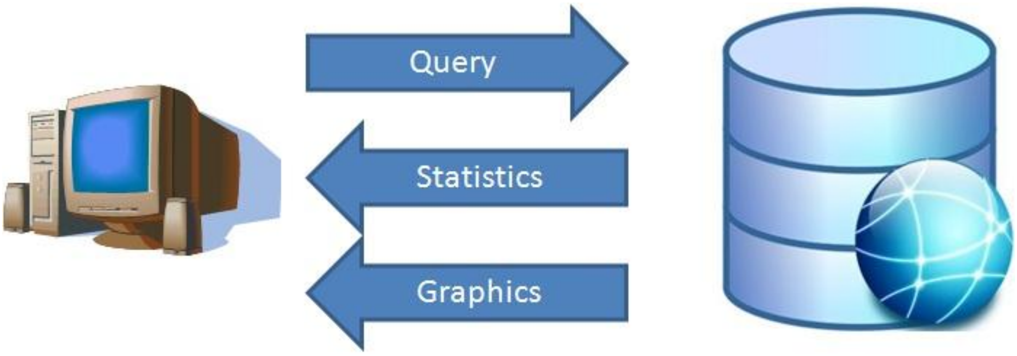


Figure 3: Automated statistical data analysis

Results are presented in reports, using tables and charts. In particular, histograms are used to show the frequency distribution of the data. Especially when data are qualitative (e. g. companies' or customers' names), they may be represented in order of decreasing frequency, in a so-called Pareto chart, in order to distinguish the "vital few" from the "trivial many", since, according to the Pareto principle, most of the effects come from a small amount of causes. For example, most of the revenues usually come from a small number of customers.

Data warehousing

Usually, in a transactional database, data are stored in flat, two-dimensional, tables. This structure allows for basic data analysis. For more sophisticated analyses, where more variables are cross-evaluated, a *data warehouse*, that is a multi-dimensional data storage facility, is needed[1].

A data warehouse is based on a dimensional model, where the main variable to be analyzed is put inside the *fact table*. If the main business process to be analyzed is sales, then the fact table contains the measures which are necessary to describe and define sales, like number of purchase orders and total sales amount.

Dimensions are then added, linking other tables by means of "join" functions, thus producing a *star schema*. Dimensions are meant to allow analysis along multiple variables. Usually, at least time and space variables are added, as in the example in figure 4, so that sales - in this case - can be analyzed in terms of historic and geographic distribution. In fact, business analysts and managers may be interested in knowing where they are performing better, if the sales trend is increasing or decreasing in the last months or years, in which period of the year sales are usually at their peak, which sales office is selling the most etc.

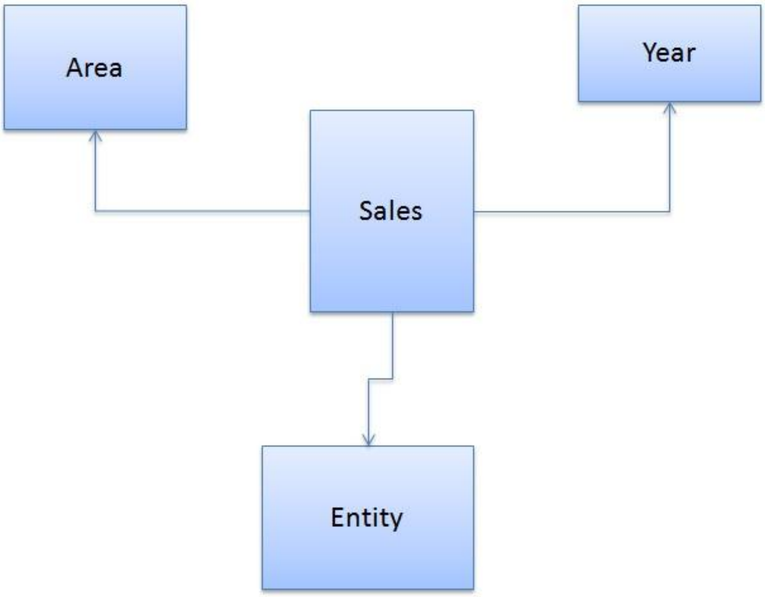


Figure 4: Star schemas (example)

Extracting data from different transactional databases and storing them inside a data warehouse is a process described as Extract, Transform and Load (ETL, see figure 5). The first part of the process, Extraction, consists in reading the data from the different available

[1] On this subject, please see [4].

data sources and copying them to the storage area of the data warehouse. The second step, Transformation, consists of data manipulations, like correcting misspellings and formats, resolving conflicts, cleansing the data etc. The storage area is accessible only to skilled professional. The final users access the presentation area, where the data are loaded, after extraction and transformation. This process leads to hyper-cubes, data warehouse structures (the equivalent of database tables), where multiple dimensions are present and can be queried and analyzed.

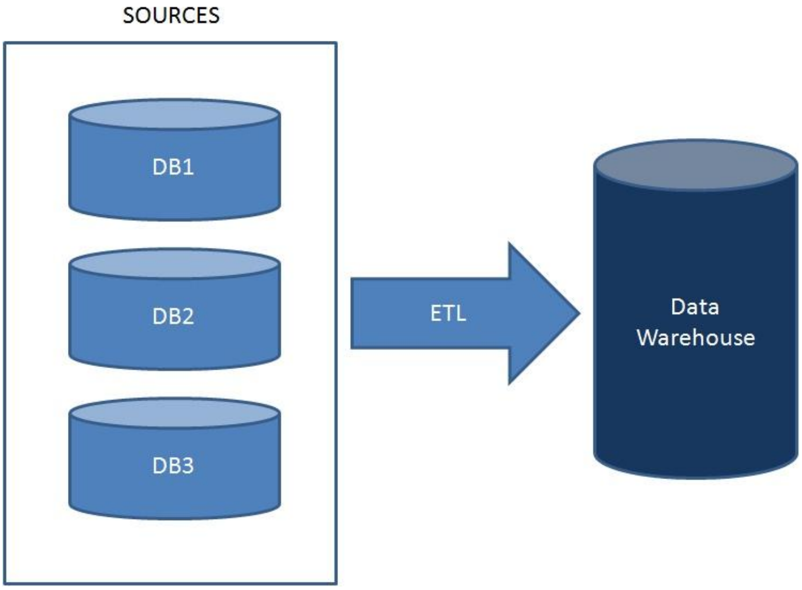


Figure 5: Extract, Transform and Load

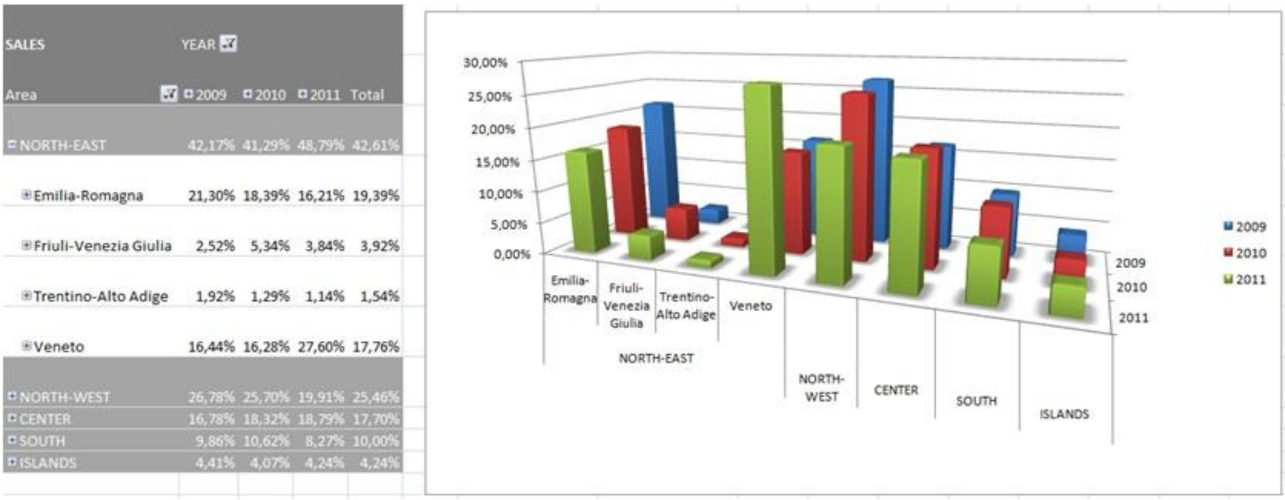
Data inside a data warehouse are time-stamped. This feature provides non-volatile data for subsequent analysis. Therefore, data warehouses can be updated periodically, but they do not change unwillingly: the database administrator must actively perform the update, loading the transaction data from the production database.

On-line Analytical Processing

On-line Analytical Processing is a set of software technologies allowing multi-dimensional queries in data warehouses. Using OLAP tools, different kind of analysis can be performed:

a. Drill-down: going into higher level of detail, opposite to consolidation. For example, we may know the global sales in the last years, but we want to know the sales per geographic area (see figure 5).
b. Roll-up (or consolidation): aggregation of data at lower levels of details, in order to have a global picture. For example, all sales offices can be grouped at a regional level and daily sales at a monthly level, thus showing regional sales per month. Eventually, means and other statistics can be calculated for time intervals or geographic areas.
c. Slicing: extracting a portion of the data from a hypercube according to a given filter (e.g. only data regarding a certain area, time period etc.).
d. Dicing: analyzing data along different dimensions (variables). For examples, we may want to know whether sales variability is associated to sales offices, period of the year or clients’ type of organization.

Results must be presented in a user-friendly format, eventually adopting traditional spreadsheet environments, like MS Excel Power Pivots. These tools allow charts to be shown beside the pivot table containing the data. The charts change accordingly to the performed function: they expand in case of drill-down or collapse in case of roll-up.



SALES	YEAR			
Area	2009	2010	2011	Total
NORTH-EAST	42,17%	41,29%	48,79%	42,61%
Emilia-Romagna	21,30%	18,39%	16,21%	19,39%
Friuli-Venezia Giulia	2,52%	5,34%	3,84%	3,92%
Trentino-Alto Adige	1,92%	1,29%	1,14%	1,54%
Veneto	16,44%	16,28%	27,60%	17,76%
NORTH-WEST	26,78%	25,70%	19,91%	25,46%
CENTER	16,78%	18,32%	18,79%	17,70%
SOUTH	9,86%	10,62%	8,27%	10,00%
ISLANDS	4,41%	4,07%	4,24%	4,24%

Figure 5: Drill-down (example)

Data mining

Data mining is a set of powerful and advanced statistical techniques used to extract meaningful, but often unpredicted, information from big data repositories[2]. Its aim is to find hidden patterns and models (associations, correlations etc.), together with counter-intuitive information. Data mining can be applied to data bases or data warehouses, in order to discover knowledge without a previously formulated hypothesis.

Often, data mining is used to understand behavior, especially customers' behavior, answering some (or more) of the following questions:

- Is there a recognizable pattern in which my customers buy goods or services?
- What classes of customers do I have?
- How can I sell more to my existing customers?

Several different data mining techniques can be successfully employed in order to get the answers. They can be divided into two broad categories: supervised and unsupervised. Supervised algorithms need inputs (other than the data) to work properly. Unsupervised algorithms instead do not need any input other then the data themselves. The following are some of the most popular data mining techniques:

a. *Cluster analysis*: Clustering is an example of unsupervised data mining and can be applied to corporate databases in order to find groups of clients with similar purchase behavior or other common characteristics, like cultural interests, income etc. Elements within a cluster are close (that is similar) to each other, while clusters are far from each other, according to the principle of strong cohesion and weak coupling.

[2] On this subject, please see [3]

For example there may be a strong but non-obvious correlation between clients coming from different geographic areas or having a similar income and their purchase or eating habits. Customer cards may be used to archive information about purchase preferences and individual data, like profession and home address. Clustering will eventually provide meaningful correlations between these data, therefore offering strategic information to the marketing staff.

b. *Link analysis*: Link or basket analysis is mainly used to discover associations between purchased items. In the case of a bank, for example, the request to convert a single account to a joint one may indicate marriage. Therefore, just-in-time advertising can improve the chances of the bank to sell a mortgage. Focusing on consumers as groups of individuals with their own habits has led to one-to-one marketing, which has proved to be more effective than the traditional one.

A supermarket chain may instead be interested in what kind of products customers tend to buy after they have bought a first one. If the management knows that those who bought a shirt may also be interested in buying a tie, then they may put them close together in the same aisle, or even put one of the two on sale. Association rules usually take the following form: ‘76% of the customers who bought A then buy B’. It is not always easy to understand the reasons behind associations, but it is easy to take advantage from them.

c. *Predictive modeling:* This technique is a kind of statistical analysis which exploits historical data to predict future trends and behavior patterns, especially to give support to marketing and Customer Relationship Management (CRM). The basic idea is that the future is contained in the past.

One of the most common techniques used in predictive modeling is the decision tree. This algorithm uses previously collected data (the *training set*) to make predictions. This process is described as *supervised learning*. The training set must contain complete and valid observations. The algorithm searches for a set of rules, which will be used to make inferences on new data. Usually, the rules can be represented as a set of “if...then” statements. Using these rules, the decision tree produces a classification, inserting every record in the database in a predefined class.

One possible application of predictive modeling is to measure attrition (that is, loss of customers). In this case, the decision tree is a simple binary tree and the possible classes into which customers will be divided are just two: stay or leave. For example, a company may predict from previously collected data, that if a customer has been with the company for longer than 2 years, then it is likely to stay. If it has not, then it is likely to leave, unless it has more than 3 services from the company (see figure 9).

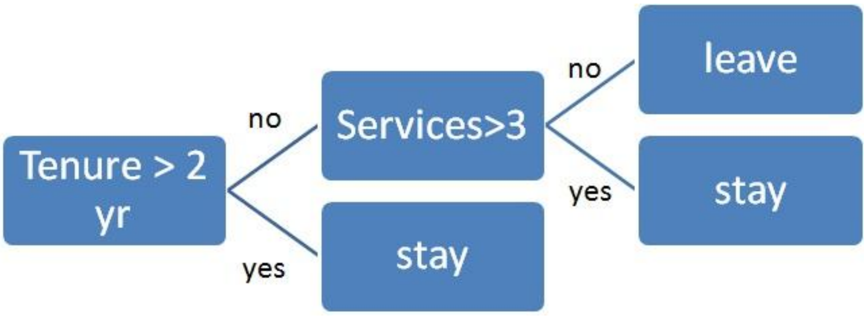


Figure 9: Attrition assessment

Business Intelligence in the public administration

Business intelligence and data mining in particular are widely used in both the private and public sectors. In the public sector, most of the applications are meant to prevent fraud or crime and reduce inefficiency.

Unfair or criminal behaviors may be spotted by means of deviation detection: if there are outliers or a skewness in a supposed-to-be normal distribution, then it is unlikely due to chance. Confronting suppliers providing the same service, data mining can spot who is eventually overcharging. Universities may use data mining applied to historical data in order to predict the percentage of students which will quit their course. Prediction is based on the students' profile.

Recently, many departments and law enforcement agencies have started to use Business Intelligence systems. For example, police may use data mining to assess crime patterns and then relocate resources accordingly. Link analysis can provide relationships between areas, type of crime and profile of the criminals (age, gender, social background, education etc.). Intelligence agencies may use data mining to look for patterns that are related to terrorist activities. Armed forces may use data mining to map and cluster their losses in different areas and historical periods, eventually predicting future losses on a statistical basis.

Performance evaluation

After issuing the law nr. 150/2009 (the so-called *Brunetta* reform, from the surname of the minister who proposed it), the Italian administration was forced to adopt a Business Intelligence system to measure its efficiency and internal production capabilities. The system relies on the measurability of the achieved results, comparing them to the expected output, in order to impose a "management by objectives" approach to public administration managers.

The main aim of the system is to reduce inefficiency. Each department of the administration is awarded a score, based on their performances. Lower-than-expected efficiency may result in increased auditing and eventual penalization. Good or excellent performances may instead be rewarded [1].

Table 1: Performance evaluation

Range	Assessment
0% - 49%	insufficient
50% - 74%	sufficient
75% - 94%	good
95% - 100%	excellent

Local administrations or ministries may use their budget in a more or less efficient way. For example, if some departments are given money to organize foreign language courses and there is a linear relationship between the amount of money and the number of organized courses, whoever falls under this distribution can be considered suspicious. Why is their performance so low? Lack of competitive suppliers? Poor management? Fraud?

Fraud detection

Fraud detection can take two different paths: fraud prevention (*a priori* fraud detection) or planning audit strategies (*a posteriori* fraud detection).

Companies asking for tax credit may be investigated *a priori* in order to see if they are likely to have the right for it or not. The probability is based on the analysis of the variables affecting such right: Are the operations actually based in an underdeveloped area? Is their activity actually technologically innovative?. Data mining evaluates the criteria that should be met by the companies to rightfully have access to the credit.

The Italian Revenue Service has adopted a profiling system to measure the average expected income of companies. Profiling is based on a set of parameters, like number of employees, sector of activity, sales, geographic area etc. Companies belonging to the same profile (that is, "class", in data mining terms) should declare an income within a narrow range around the mean. Whenever the declared income falls below the average, a warning signal is raised and auditing is enforced. If instead the declared income is above the average, auditing is reduced. Since the amount of resources which can be devoted to auditing is limited, better performances are achieved when auditing is enforced for companies which fall below the line, where there is a greater probability of finding fraud.

Data mining can determine the voluntary compliance of tax payers by classifying or clustering them according to historical data. Tax payers belonging to each cluster will then be awarded a score of compliance, indicating the likelihood that they are willing to pay taxes, regardless of eventual improvements in enforcement. Clusters have different values for different variables and are therefore peculiar in terms of job, age, education, geographic distribution etc. Enforcement will be increased for the clusters with low scores.

Several studies have been devoted to automated fraud detection [6]. Data mining in particular has been used to optimize audit strategies, in case of tax evasion [2, 7] or financial fraudulent statements [5], proving to be an effective technique to fight fraud and crime.

Conclusions

Business Intelligence can be effectively implemented in both the private and public sectors. In private companies, it is mainly used to improve their competitive advantage, by means of market analysis, basket analysis and attrition assessment.

In the public administration, BI can be used to measure performances and improve efficiency. In law enforcement, data mining can be used to prevent crime and optimize resources. In particular, predictive modeling and classification have proved to be effective in preventing or detecting tax evasion and fraudulent financial statements.

References

[1] Il bilancio dell'esercito e il controllo interno di gestione. Rapporto Esercito 2010, pages 120-131, 2011. Supplemento alla Rivista Militare

[2] F. Bonchi, F. Giannotti, G. Mainetto, and D. Pedreschi. Using data mining techniques in fiscal fraud detection. In DaWak'99, First Int. Conf. on Data Warehousing and Knowledge Discovery, 1999

[3] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi. *Discovering Data Mining*. Prentice Hall, Upper Saddle River, NJ (USA), 1998

[4] R. Kimball and M. Ross. *The Data Warehouse Toolkit*. Wiley, New York (USA), 2nd edition, 2002

[5] E. Kirkos, C. Spathis, and Y. Manolopulos. Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 32:995-1003, 2007

[6] Clifton Phua, Vincent Lee, Kate Smith-Miles, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. 2005

[7] Fan Yu, Zheng Qin, and Xiao-Ling Jia. Data mining application issues in fraudulent tax declaration detection. In Machine Learning and Cybernetics, International Conference on, vol. 4, pages. 2202-2206, Nov. 2003