THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

DATABASE SYSTEMS JOURNAL

Vol. IV, Issue 4/2013

LISTED IN

RePEc, EBSCO, DOAJ, Open J-Gate,
Cabell's Directories of Publishing Opportunities,
Index Copernicus, Google Scholar

BUSINESS INTELLIGENCE

ERP

DATA MINING

DATA WAREHOUSE

DATABASE

ISSN: 2069 – 3230

Contact
Calea Dorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: http://dbjournal.ro
E-mail: editordbjournal@gmail.com; editor@dbjournal.ro

CONTENTS

E-COCOMO: The Extended COst Constructive MOdel for Cleanroom Software Engineering

Hitesh KUMAR SHARMA
University of Petroleum and Energy Studies, India
hkshitesh@gmail.com

Mistakes create rework. Rework takes time and increases costs. The traditional software engineering methodology defines the ratio of Design:Code:Test as 40:20:40. As we can easily see that 40% time and efforts are used in testing phase in traditional approach, that means we have to perform rework again if we found some bugs in testing phase. This rework is being performed after Design and code phase. This rework will increase the cost exponentially. The cleanroom software engineering methodology controls the exponential growth in cost by removing this rework. It says that "do the work correct in first attempt and move to next phase after getting the proof of correctness".
This new approach minimized the rework and reduces the cost in the exponential ratio. Due to the removal of testing phase, the COCOMO (COst COnstructive MOdel) used for the traditional engineering is not directly applicable in cleanroom software engineering. The traditional cost drivers used for traditional COCOMO needs to be revised. We have proposed the Extended version of COCOMO (i.e. E-COCOMO) in which we have incorporated some new cost drivers. This paper explains the proposed E-COCOMO and the detailed description of proposed new cost driver.
Keywords: *Cleanroom Software Engineering, COCOMO, Effort Estimation, Cost Drivers, SDLC.*

1 Introduction

Harlan Mills and his colleagues from IBM developed the CSE (Cleanroom Software Engineering) methodology in the early 1980s. They were part of IBM's Federal Defense System where software failures could mean millions of dollars and most importantly, human lives. In This software methodology they used the same analogy as used in cleanroom fabrication of semiconductors. Instead of trying to clean dirt off the semiconductor wafers after production, the object is to prevent the dirt from getting into the production environment in the first place. The reason for this is that defect prevention is more cost effective than defect removal. Therefore, in software development, the CSE methodology eliminates or avoids as many defects as possible before software execution using controlled and measurable statistics. Because of that reason they start the cleanroom software Development methodology for software development.

The **Constructive Cost Model** (**COCOMO**) is an algorithmic software cost estimation model developed by Barry Boehm. The model uses a basic regression formula, with parameters that are derived from historical project data and current project characteristics. COCOMO was first published in 1981 Barry W. Boehm's Book *Software engineering economic* as a model for estimating effort, cost, and schedule for software projects. It drew on a study of 63 projects at TRW Aerospace where Barry Boehm was Director of Software Research and Technology in 1981. The study examined projects ranging in size from 2,000 to 100,000 lines of code, and programming languages ranging from assembly to PL/I. These projects were based on the waterfall model of software

development which was the prevalent software development process in 1981.

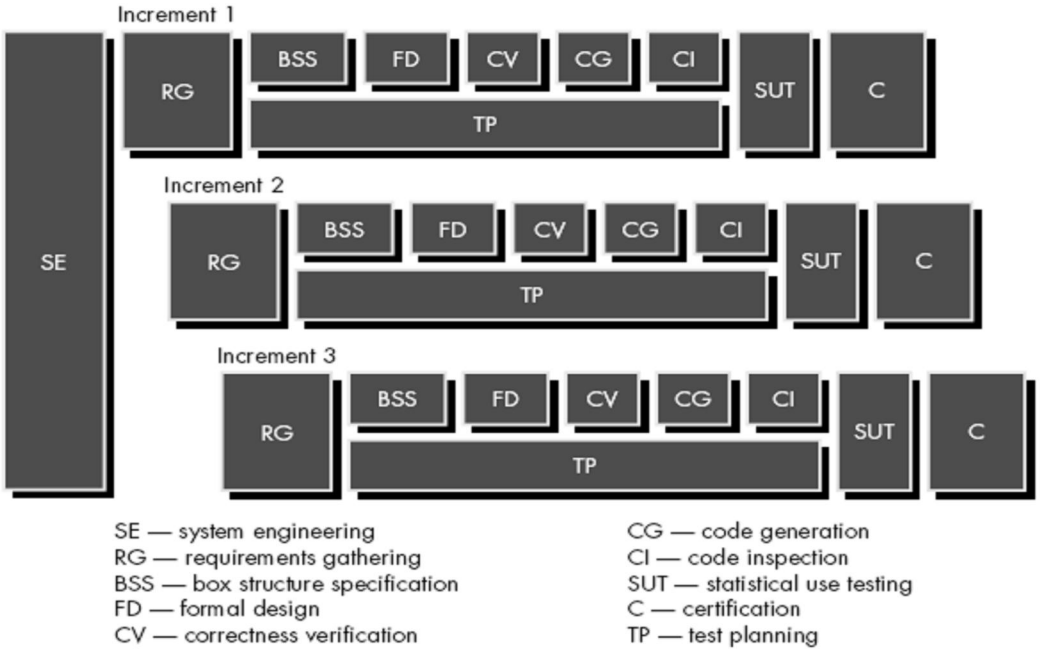
2. Cleanroom Software Engineering (CSE)

The cleanroom approach makes use of a specialized version of the incremental software model. A "pipeline of software increments" is developed by small independent software engineering teams. As each increment is certified, it is integrated in the whole. Hence, functionality of the system grows with time. The sequence of cleanroom tasks for each increment is illustrated in Figure 1. Overall system or product requirements are developed using the system engineering methods. Once functionality has been assigned to the software element of the system, the pipeline of cleanroom increments is initiated. The following tasks occur in CSE:

Increment planning. A project plan that adopts the incremental strategy is developed. The functionality of each increment, its projected size, and a cleanroom development schedule are created. Special care must be taken to ensure that certified increments will be integrated in a timely manner.

Requirements gathering. Using traditional techniques, a more-detailed description of customer-level requirements (for each increment) is developed.

Box structure specification. A specification method that makes use of box structures is used to describe the functional specification. Box structures "isolate and separate the creative definition of behavior, data, and procedures at each level of refinement."



Formal design. Using the box structure approach, cleanroom design is a natural and seamless extension of specification. Although it is possible to make a clear distinction between the two activities, specifications (called *black boxes*) are iteratively refined (within an increment) to become analogous to architectural and component-level designs (called *state boxes* and *clear boxes,* respectively).

Correctness verification. The cleanroom team conducts a series of rigorous

correctness verification activities on the design and then the code. Verification begins with the highest-level box structure (specification) and moves toward design detail and code. The first level of correctness verification occurs by applying a set of "correctness questions". If these do not demonstrate that the specification is correct, more formal (mathematical) methods for verification are used.

Code generation, inspection, and verification. The box structure specifications, represented in a specialized language, are translated into the appropriate programming language. Standard walkthrough or inspection techniques are then used to ensure semantic conformance of the code and box structures and syntactic correctness of the code. Then correctness verification is conducted for the source code.

Statistical test planning. The projected usage of the software is analyzed and a suite of test cases that exercise a "probability distribution" of usage are planned and designed Referring to Figure 1, this cleanroom activity is conducted in parallel with specification, verification, and code generation.

Statistical use testing. Recalling that exhaustive testing of computer software is impossible, it is always necessary to design a finite number of test cases. Statistical use techniques execute a series of tests derived from a statistical sample (the probability distribution noted earlier) of all possible program executions by all users from a targeted population.

Certification. Once verification, inspection, and usage testing have been completed (and all errors are corrected), the increment is certified as ready for integration. Like other software process models discussed elsewhere in this book, the cleanroom process relies heavily on the need to produce high-quality analysis and design models. As we will see later in this chapter, box structure notation is simply another way for a software engineer to represent requirements and design. The real distinction of the cleanroom approach is that formal verification is applied to engineering models.

Dyer alludes to the differences of the cleanroom approach when he defines the process:

"*Cleanroom represents the first practical attempt at putting the software development process under statistical quality control with a well-defined strategy for continuous process improvement. To reach this goal, a cleanroom unique life cycle was defined which focused on mathematics based software engineering for correct software designs and on statistics-based software testing for certification of software reliability."*

Cleanroom software engineering differs from the conventional and object-oriented views because:

- It makes explicit use of statistical quality control.
- It verifies design specification using a mathematically based proof of correctness.
- It relies heavily on statistical use testing to uncover high-impact errors.

Obviously, the cleanroom approach applies most, if not all, of the basic software engineering principles and concept. Good analysis and design procedures are essential if high quality is to result. But cleanroom engineering diverges from conventional software practices by deemphasizing (some would say, eliminating) the role of unit testing and debugging and dramatically reducing (or eliminating) the amount of testing performed by the developer of the software. In conventional software development, errors are accepted as a fact of life. Because errors are deemed to be inevitable, each program module should be unit tested (to uncover errors) and then

debugged (to remove errors). When the software is finally released, field use uncovers still more defects and another test and debug cycle begins. The rework associated with these activities is costly and time consuming. Worse, it can be degenerative error correction can (inadvertently) lead to the introduction of still more errors. In cleanroom software engineering, unit testing and debugging are replaced by correctness verification and statistically based testing. These activities, coupled with the record keeping necessary for continuous improvement, make the cleanroom approach unique.

3. Formal specification

Formal methods allow a software engineer to create a specification that is more complete, consistent, and unambiguous than those produced using conventional or object oriented methods. Set theory and logic notation are used to create a clear statement of facts (requirements). This mathematical specification can then be analyzed to prove correctness and consistency. Because the specification is created using mathematical notation, it is inherently less ambiguous than informal modes of representation. A specially trained software engineer creates a formal specification. In safety-critical or mission critical systems, failure can have a high price. Lives may be lost or severe economic consequences can arise when computer software fails. In such situations, it is essential that errors are uncovered before software is put into operation. Formal methods reduce specification errors dramatically and, as a consequence, serve as the basis for software that has very few errors once the customer begins using it. The first step in the application of formal methods is to define the data invariant, state, and operations for a system function. The data invariant is a condition that is true throughout the execution of a function that contains a collection of data, The state is the stored data that a function accesses and alters; and operations are actions that take place in a system as it reads or writes data to a state. An operation is associated with two conditions: a precondition and a post condition. The notation and heuristics of sets and constructive specification set operators, logic operators, and sequences form the basis of formal methods. A specification represented in a formal language such as Z or VDM is produced when formal methods are applied.

4. COCOMO (COnstructive COst MOdel)

Boehm's COCOMO model is one of the mostly used model commercially. The first version of the model delivered in 1981 and COCOMO II is available now. COCOMO'81 is derived from the analysis of 63 software projects in 1981. Boehm proposed three levels of the model :

- Basic COCOMO
- Intermediate COCOMO
- Detailed COCOMO

4.1 Basic COCOMO

Basic COCOMO computes software development effort (and cost) as a function of program size. Program size is expressed in estimated thousands of lines of code (KLOC). COCOMO applies to three classes of software projects:

Table 1. Classes of Projects

Project Class	Project Size	Nature of Project	Deadline	Development Environment
Organic	Typically 2-50 KLOC	Small size project, experienced developers in the familiar environment. For example, pay roll, inventory projects etc.	Not tight	Simple/Familiar/ In-house
Semi-Detached	Typically 50-300 KLOC	Medium size project, Medium size team, Average previous experience on similar project. For example: Utility systems like compilers, database systems, editors etc.	Medium	Medium
Embedded	Typically over 300 KLOC	Large project, Real time systems, Complex interfaces, Very little previous experience. For example: ATMs, Air Traffic Control etc.	Tight	Complex

Formula for Basic COCOMO

$$E = a_b(KLOC)^{b_b}$$

$$D = c_b(E)^{d_b}$$

where E is effort applied in Person-Months, and D is the development time in months. The coefficients a_b, b_b, c_b and d_b are given in table 2:

Table 2. Coefficients a_b, b_b, c_b and d_b values

Software Project	a_b	b_b	c_b	d_b
Organic	2.4	1.05	2.5	0.38
Semidetached	3.0	1.12	2.5	0.35
Embedded	3.6	1.20	2.5	0.32

Basic COCOMO is good for quick estimate of software costs. However it does not account for differences in hardware constraints, personnel quality and experience, use of modern tools and techniques, and so on.

4.2 Intermediate COCOMO

Intermediate COCOMO computes software development effort as function of program size and a set of "cost drivers" that include subjective assessment of product, hardware, personnel and project attributes. This extension considers a set of four "cost drivers", each with a number of subsidiary attributes:

- Product attributes
  - Required software reliability
  - Size of application database
  - Complexity of the product
- Hardware attributes
  - Run-time performance constraints

  - Memory constraints
  - Volatility of the virtual machine environment
  - Required turn about time
- Personnel attributes
  - Analyst capability
  - Software engineering capability
  - Applications experience
  - Virtual machine experience
  - Programming language experience
- Project attributes
  - Use of software tools
  - Application of software engineering methods
  - Required development schedule

Each of the 15 attributes receives a rating on a six-point scale that ranges from "very low" to "extra high" (in importance or value). An effort multiplier from the table below applies to the rating. The product of all effort multipliers results in an *effort adjustment factor (EAF)*. Typical values for EAF range from 0.9 to 1.4.

The Intermediate COCOMO formula now takes the form:

$$E = a_i (KLOC)^{b_i} * EAF$$
$$D = c_i (E)^{d_i}$$

where E is the effort applied in person-months, **KLoC** is the estimated number of thousands of delivered lines of code for the project, and **EAF** is the factor calculated above. The coefficient $\mathbf{a_i}$ and the exponent $\mathbf{b_i}$ are given in the next table.

Table 3. Coefficients a_i, b_i, c_i and d_i values

Project	a_i	b_i	c_i	d_i
Organic	3.2	1.05	2.5	0.38
Semidetached	3.0	1.12	2.5	0.35
Embedded	2.8	1.20	2.5	0.32

The Development time **D** calculation uses **E** in the same way as in the Basic COCOMO.

4.3 Detailed COCOMO

Detailed COCOMO is defined in Barry Boehm's book "Software Engineering Economics in 1981". Detailed COCOMO incorporates all characteristics of the Intermediate COCOMO version with an assessment of the cost driver's impact on each step (analysis, design, etc.) of the software engineering process. Detailed COCOMO offers a means for processing all the project characteristics to construct a software estimate. The detailed model introduces two more capabilities:

The formula for detailed COCOMO is:

$$E_p = \mu_p E$$
$$D_p = \tau_p D$$

5. E-COCOMO (Extended COst COnstructive MOdel)

As we have discussed in intermediate COCOMO that there are 15 cost driver factors in traditional software engineering to calculate EAF. But as we are moving towards Cleanroom methodology in software development we need some new cost drivers which will be incorporated due to the inclusion of BSS and Formal Specification. The drivers should be added to the personal attribute category because the humans involve in CSE process should have the knowledge of these new included components. Due to the need to include

some new cost driver we found to add one new cost driver in Intermediate COCOMO that is "**Formal Method Knowledge Capability(FMKC)**". It specify the knowledge experience of Formal Method and Formal Specification Language like 'Z' Specification language. Formal Method knowledge it must require for the cleanroom Development Mythology. Formal Methods used in developing computer systems are mathematically used techniques for describing system properties. The four phases used in the detailed COCOMO model are: requirements planning and product design (RPD), detailed design (DD), code and unit test (CUT), and integration and test (IT) that is based on Waterfall model if cleanroom development mythology used then these phase will changed. We proposed to use Four phase in Detailed COCOMO model are: Increment planning and Requirement gathering (IPRG), Box structure specification and Formal Design (BSSFD), Correctness verification and code generation(CVCG),Statistical Test planning and Use Testing(STPUT).

Table 4. Table for multiplying factors for EAF
(The values for new cost driver "FMKC" is highlighted)

Cost Drivers	**Ratings**					
	Very Low	Low	Nominal	High	Very High	Extra High
Product attributes						
RELY	0.75	0.88	1	1.15	1.4	
DATA		0.94	1	1.08	1.16	
CPLX	0.7	0.85	1	1.15	1.3	1.65
Hardware attributes						
TURN			1	1.11	1.3	1.66
VIRT			1	1.06	1.21	1.56
STOR		0.87	1	1.15	1.3	
TIME		0.87	1	1.07	1.15	
Personnel attributes						
ACAP	1.46	1.19	1	0.86	0.71	
LEXP	1.29	1.13	1	0.91	0.82	
VEXP	1.42	1.17	1	0.86	0.7	
PCAP	1.21	1.1	1	0.9		
AEXP	1.14	1.07	1	0.95		
FMKC	1.43	1.18	1	0.86	0.7	-
Project attributes						
MODP	1.24	1.1	1	0.91	0.82	
SCED	1.24	1.1	1	0.91	0.83	
TOOL	1.23	1.08	1	1.04	1.1	

The values of the coefficient (i.e. Effort coefficient μ_p and Time Coefficient $\tau_{p)}$ used it will also change in Detailed COCOMO. The modified values have been shown in the following tables.

Table 5. Table for E-COOCMO μ_p used for cleanroom engineering phases

Mode & code size	IRPG	BSSFD	CVCG	STPUT
Organic small	0.15	0.65	0.17	0.03
Organic medium	0.15	0.64	0.17	0.04
Semidetached medium	0.16	0.64	0.16	0.04
Semidetached large	016	0.63	0.15	0.06
Embedded large	0.18	0.62	0.14	0.06
Embedded extra large	0.18	0.61	0.14	0.07

Table 6. Table for E-COOCMO τ_p used for cleanroom engineering phases

Mode & code size	IRPG	BSSFD	CVCG	STPUT
Organic small	0.14	0.66	0.17	0.03
Organic Medium	0.14	0.65	0.17	0.04
Semidetached Medium	0.15	0.65	0.16	0.04
Semidetached Large	015	0.64	0.15	0.06
Embedded Large	0.17	0.63	0.14	0.06
Embedded extra large	0.17	0.62	0.14	0.07

Conclusion & future work

The software industries are adopting the new methodologies and leaving the traditional methodology far behind. Due to this transition the metrics and measurement based on the traditional methodology should also change. The old methods for effort and time calculation cannot apply on new development methodologies. To keep this transition in mind we have defined some new parameter those should be included in traditional COCOMO to calculate the effort and time for a software project. We have given a new name to this new version of COCOMO as E-COCOMO (i.e. Extended COst COnstructive MOdel). This model can be used to calculate effort and time for the projects those are adapting cleanroom software engineering methodology.

In future the work will be extended for other development methodologies (i.e Agile Development, Object Oriented Development, Component Based Engineering etc.). These methodologies cannot use traditional COCOMO to calculate the efforts and time. Some enhancement is need in traditional COCOMO to calculate exact results.

References

[1] M. Wolak , Taking the Art out of Software Development. An In-Depth Review of Cleanroom software Engineering by Chaelynne.

[2] Linger, R.C., "Cleanroom Process Model," IEEE Software. March 1994, pp. 50–58.

[3] Hevner, A.R. and H.D. Mills, “Box Structure Methods for System Development with Objects,” IBM Systems Journal, vol. 31, no.2, February 1993, pp. 232–251.

[4] Linger, R.M. and H.D. Mills, “A Case Study in Cleanroom Software Engineering: The IBM COBOL Structuring Facility,” Proc. COMPSAC ’88, Chicago, October 1988.

[5] Poore, J.H. and H.D. Mills, "Bringing Software Under Statistical Quality Control," Quality Progress, November 1988, pp. 52–55.
[6] Dyer, M., The Cleanroom Approach to Quality Software Development, Wiley, 1992.
[7] Harlan D. Mills, Michael Dyer and Richard C. Linger, "Cleanroom Software Engineering", IEEE software, September 1987.
[8] Robert Oshana and Frank P. Clyle "Implementing cleanroom Software engineering into a mature CMM-based software organization" Proceedings of the 1997 International Conference on Software Engineering, Boston United States, pp: 572-573, May 1997.
[9] Richard C. Linger "Cleanroom Software engineering for zero-defect software",Proceedings of the 15th international conference on software engineering, Baltimore.
[10] Boehm, B.W. (1981). Software Engineering Economics. Prentice Hall.
[11] K. K. Agarwal. "Software Engineering".
[12] R. Pressman , "A practioner approach for software engineering". Fifth Edition.
[13] Mills, H.D., M. Dyer, and R. Linger, "Cleanroom Software Engineering," IEEE Software, vol. 4, no. 5, September 1987, pp. 19–24.
[14] Wohlin, C. and P. Runeson, "Certification of Software Components," IEEE Trans. Software Engineering, vol. SE-20, no. 6, June 1994, pp. 494–499.
[15] Hausler, P.A., R. Linger, and C. Trammel, "Adopting Cleanroom Software Engineering with a Phased Approach," IBM Systems Journal, vol. 33, no.1, January 1994, pp. 89–109.

Hitesh KUMAR SHARMA is an Assistant Professor in University of Petroleum & Energy Studies, Dehradun. He has published 8 research papers in National Journals and 5 research papers in International Journal. Currently He is pursuing his Ph.D. in the area of database tuning**.**

Business Intelligence Systems

Bogdan NEDELCU
University of Economic Studies, Bucharest, Romania
bogdannedelcu@hotmail.com

The aim of this article is to show the importance of business intelligence and its growing influence. It also shows when the concept of business intelligence was used for the first time and how it evolved over time. The paper discusses the utility of a business intelligence system in any organization and its contribution to daily activities. Furthermore, we highlight the role and the objectives of business intelligence systems inside an organization and the needs to grow the incomes and reduce the costs, to manage the complexity of the business environment and to cut IT costs so that the organization survives in the current competitive climate. The article contains information about architectural principles of a business intelligence system and how such a system can be achieved.
Keywords: *Business Intelligence, Data warehouse, OLAP*

1 The Business Intelligence Concept and its Appearance

The concept of business intelligence became more and more used during the last years, and now, this association of terms is used across different fields from data technology to business modeling.
Business intelligence represents a wide area of applications and technologies for collecting, storing, analyzing and providing access to information for improving businesses process modeling quality. [1]
The business intelligence statement, „getting the right information to the right people at the right time" [2] focuses on the fact that business intelligence uses information and not data due to the included capabilities for processing raw data into intelligent information, that is valid and accepted by the entire company and which can be consistently used in process modeling.
The business intelligence term was introduced by Gartner Group in middle of the 90s. The concept, on the other hand, existed long before being used in mainframe reporting systems.
The Gartner Group defines business intelligence as "an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance". [2]
It was due to reporting and analyzing requirements that the need for such systems keeps the same growth rate as technology does.
We can all agree that in the current economic environment, information and its accuracy represent a successful key to any business. Though it's about accomplishing a simple sales report or extracting raw data from a database, analyzing data always provided precious information to company managers.
Over time, the interesting data was provided to persons with decisional role within a company as reports, statistics or analyzing documents. But what happens when the volume of data increases so much that classical storage and reporting systems can't keep up with it?
At the beginning, some decided to archive old data and keep the current data. This solution worked perfectly until it was realized that old data is as good as the new

one and then in started the reversed process. Unfortunately, all these steps wore only complicating the process of analyzing and elaborating reports.

A solution was necessary in that moment. Big companies realized that they can't keep up with their competitors. More and more complex products wore appearing on the market. Exploiting precious resources like a company's historical data needed to be approached with a new technique. It was just a matter of time until the conceptual answer and the technical one appeared, and that answer was business intelligence systems.

The business intelligence term was used for the first time in 1958, when Hans Peter Luhn used it in an article. Business intelligence, as it is known today, it's supposed to be developed form decision support systems which evolution begun in the late 60s, reaching a major point in the middle 80s. Decision support systems had their origin within computer assisted models which wore created with the purpose of helping company management in taking decisions and performance scheduling. From these decision support systems took birth in the late 80s the concepts of data warehouse, OLAP (on-Line Analytical Processing) and business intelligence. [3]

In 1989, Howard Dresner used the term of business intelligence to group under the same topic all methods and techniques used for taking decisions which wore based on solid facts. Starting with the 90s, the frequency of using the business intelligence term has increased even more.

The 1990s represented a formative period for business intelligence applications and products [4]. During this period, organizations realized that they needed arrange of business intelligence capabilities to satisfy a diverse set of user needs. Ultimately, this range could be grouped into five distinct categories or "Styles of Business Intelligence" applications. The 5 styles of Business Intelligence applications include:

a) Data Mining and Advanced Analysis
b) Visual and OLAP analysis
c) Enterprise Reporting
d) Dashboards and Scorecards
e) Mobile Apps and Alerts

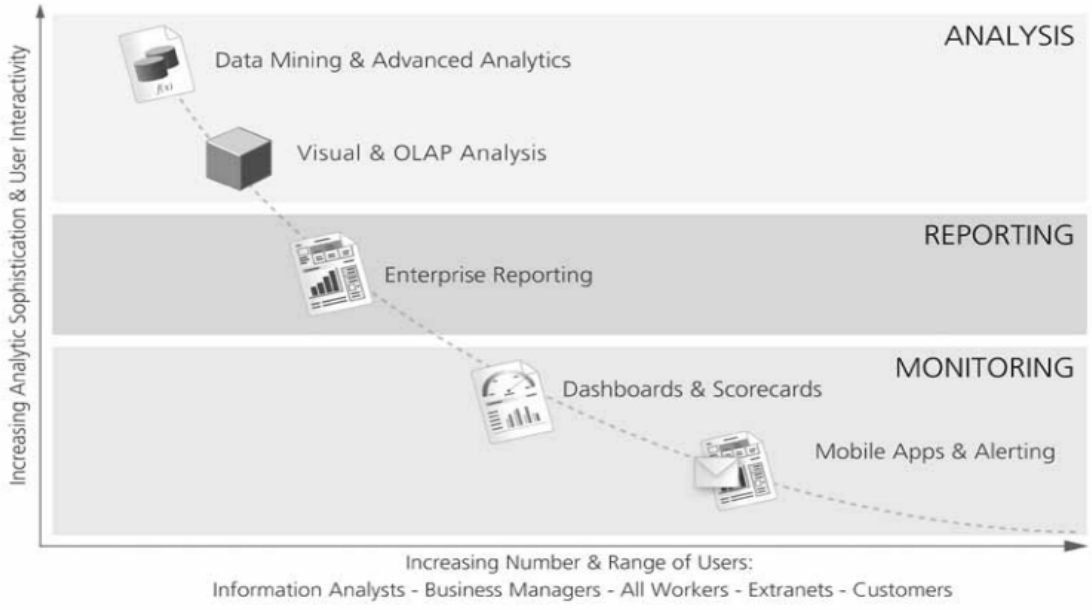


Fig. 1. The Microstrategy architecture that offers all 5 Style of Business Intelligence with a single organically-developed architecture [4]

Nowadays, investments have risen in a field like business intelligence. Software developers haven't delayed in developing special solutions for business intelligence, so big processing machines wore made for using such systems and the research in the field grew. Business intelligence is strictly tied to technologies like data mining, OLAP, data connections and data warehouses.

Therefore, business intelligence is based on an aggregate of concepts and technologies which cooperate for helping companies and their decisional activities. As we said before, usually, an organization must own at least one storing and processing data system.

In the current environment, everything is based on information; companies provide informational activities needs and the internet makes the information transfer without having distances barriers.

Everything is information. Companies get to own, as we said before, data that can no longer be stored by the classic database systems. Researches in business intelligence revealed that storing, analyzing and exploiting an organizations data could provide precious information to a company such as predictions, patterns or complex reports.

2. The Utility of a Business Intelligence System

Every organization has a number of informational systems that contributes to its good ongoing daily activities. Though it's about an informatics system for client management, sales, payments, or human resources, all these are managing the daily process transactions, and are being designed for that purpose. Usually, these systems are known as transactional systems (OLTP – On-line transactional processing), and their essential role is to assure a good data consistency.

Most transactional systems have an infrastructure based on a relational data base specially designed for the well function of these systems. The data model on which this systems are based is represented in the entity-relation diagram, which leads to a normalized structure, in which the information regarding a certain entity are stored in a single table, and the link between entities is made with relations. The relational model, with the principles that it implements, proved to be a very good solution for transactional systems, being implemented and used, even currently, by the majority of organizations. This gathers information in a single place, removing the redundancy if possible, so that the data could be easily found and updated, to overcome the big number of daily transactions.

A business, no matter how big it is, doesn't assume only entering the data in a system. The data must be uploaded with a purpose, to keep track of the sales, for example, therefore most transactional systems offer complex utilities to help generating reports. Unfortunately, these reports will be limited strictly to the operational or transactional system which generated them, and the effort of understanding how the whole business works, based on these limited reports, is by far unsuccessful.

The complex questions to which analysts would like to find answers through reports usually needs data from more than one transactional system. Though the data can be extracted from a system, to have a general view of the business, the data should be correlated with data from other systems. What will generating reports based on data from different systems imply? A well knowing of the systems, the technologies in which they are built in, filtrating the unwanted data, etc. Because people with decisional role in a company are usually less technical, generating such reports would be impossible to them.

Business intelligence comes with a solution to all these.
A business intelligence system is a mix of technologies and concepts that are specially designed to help managers take decisions. How does an informatic system really help? As we said before, the data owned by an organization is an important and useful source of information. Based on these data you can generate reports, predictions, sale charts or establish the best market segment. Unfortunately, a company cannot always store its data in transactional databases, because their volume would significantly slow the data processing time. Therefore, many companies choose to implement a business intelligence system.
A business intelligence system is based, in 90% of the cases, on the implementation of a different database then the transactional one. Usually, the implemented database for such a system will have huge dimensions, being installed on machines with high processing capacity and it will be designed for massive queries. All data relevant to the business analyze will be inserted in this database. The basic data on which ideas can be stated and from which relevant information for the organization is extracted, is established in the analyzing and designing stage of a business intelligence system.
Certainly, not the entire data of an organization will be relevant for analyze. A part of the data will be filtered, precisely to provide the answers that analysts look for. Once the data that needs processing and their transactional system location is established, the database design for the business intelligence system can start. This special database, having the purpose to store the historical data, is built as a data warehouse or as a data concentration.
Developing and implementing an historical database it's not enough for providing business intelligence. This data must be exploited so that information is extracted. Major software companies have developed various tools for complex reports and dashboards. There are various methods for harvesting data (data mining), and the OLAP (On-Line Analytical Processing) technology is, usually, the mainstay of a business intelligence system.

3. The Role and Objectives of Business Intelligence Systems inside an Organization

Business intelligence solutions bring added value inside enterprises, serving as a base for making fundamental changes, new collaborations, acquiring new customers, creating new markets.

- *The business intelligence system's role in taking decisions*

At a strategic level, business intelligence makes possible the establishing of objectives in a precise mode and following their achievement, allowing different comparative reports, also performing growth simulations or forecasting the next result on the base of some assumptions.
At a tactic level, the business intelligence system can offer a base for making marketing, sales, finances, income or management decisions. The systems allow the optimization of future actions and performance modifiers on an organizational aspect, financial or technologic on the purpose to help enterprises to reach their strategic objectives in a more efficient way.
Regarding the operational level, business intelligence solutions are used in establishing the ad-hoc analyzing and answering questions related to suppliers cooperation, clients and operations already in developing inside each department.
The main objectives of the business intelligence systems are resumed shortly in the next figure:

- *Using business intelligence solutions*

In the competitive climate of our days, it's vital for organization to offer a fast access to information, at low cost, for a larger number and variety, for the users. The solution of this problem is a business intelligence system that offers a set of technologies and software products that gives the users the needed information to answer the questions that appear in solving business problems.

a. The need to grow the incomes and reduce the costs
The days where the user could manage and plan the activities using monthly reports and IT organizations had a lot of time on their disposal to implement new applications, are over. Today companies need to quickly develop applications in order to offer their users a faster and easier access to information that reflect changes of the business environment. Business intelligence systems focus on fast delivery and access of information to users

b. The need to manage the complexity of the business environment
It becomes harder to understand and manage a complex business environment and to maximize the investments. Business intelligence systems offer more than queries and reports solutions, they offer analyzing instruments of complex information and data mining.

c. The need to cut IT costs
Today, the investment in IT systems is a significant percent of companies' costs. It is not necessary just to reduce these costs, but also, to obtain maximum benefits from the information managed by the IT systems. The new IT technologies like the Intranet and 3 level architectures reduce the cost of using the business intelligence systems by a large variety of users, especially managers.

4. Architectural Principles of a Business Intelligence System

In an overview of the microstrategy platform architecture for big data [4], cloud business intelligence and mobile application, the architectural principles of a business intelligence system are stated as following:

a) Scalability and High Performance
All design decisions must ensure that the strategy can deliver consistently high performance as the system scales upward, and must anticipate order-of-magnitude growth beyond today's state-of-the-art standard (user scale, data scale and application scale).

b) Economies of Scale
All design decisions must explicitly deliver greater economies of scale as a system grows – using techniques like in-memory data processing, caching, object reuse, automated administration, and collaborative analytics. We make sure that the strategy implementations require an absolute minimum of IT personnel, require the fewest servers, and minimize the workload on expensive database resources.

c) Complete Functionality
All designs decisions must ensure that architecture offers the full range of business intelligence functionality on a single service-oriented architecture, so that customers can satisfy all of their business intelligence requirements without the need for additional integration work.

d) Incremental Growth
All design decisions must ensure that customers can incrementally grow their business intelligence infrastructures – from small to large, from departmental scope to enterprise cope, from isolated islands to consolidated applications, and from reporting to dashboards to OLAP to ad hoc

analysis to alerting to mobile apps. This allows customers to initially buy just the functionality they need, and to incrementally grow their business intelligence solution as their requirements naturally expand.

e) Openness and Extensibility
All design decisions must ensure that the strategy's vast functionality continues to be fully accessible through Web services APIs.

f) Centralized Consistency with Distributed Governance and Self-service
All design decisions must support the goal of a consistent single version of the truth throughout the enterprise using a single shared metadata and pervasive security architecture. Yet, the architecture must also provide a high degree of autonomy to distributed development teams (managed by departments and divisions), and to individual users allowing them to create enterprise-consistent solutions at a local level.

g) Rapid Development and Deployment
All design decisions must promote rapid development and deployment of new reports and applications. Some developers have invested significant engineering energy in creating a vast array of reusable metadata objects, by creating a security architecture that is applied automatically and pervasively with no effort on the part of the report designer, and through design paradigms that allow novices and experts to play a role in accelerating the report design process.

h) Consistent Experience
Developers continuously work towards providing the same business intelligence experience from any user interface– desktop, web browsers, or mobile devices. Any feature, however simple it may be, is added to the platform in a way that it can be easily available from any user interface. This philosophy enables business users to seamlessly change their interface to access critical business reports without losing any functionality. These requirements are from long-time customers who invested in high-scale business intelligence applications. Today, the goal of these same customers is to host many diverse business intelligence applications on a Cloud instance, or make these applications Mobile. The architectural tenets discussed earlier ensure their long-term success in this process.

5. Achieving a Business Intelligence System

When a company decides to implement a system for business intelligence, it's good to consider that the implementation of this kind of technology is based on a very long and laborious process. Usually, just the analyzing stage can take a year, of course, considering the size of the organization and the complexity of the business behind. Another very important aspect to mention is the one of high costs that the developing of such a system involves; according to the statistics, the amortization of developing costs of a business intelligence system is done in a few years.

a) Transactional systems as a data source
The first stage in any business intelligence system is always the analyzing stage. In this first stage, the aims of the business intelligence system are set. As we mentioned before, any organization has a mix of operational systems, also known as OLTP Systems (On Line Transactional Processing Systems). These systems are usually used to process the company's current transactions and can be used for managing clients, sales or suppliers. These can be ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), POS (Point Of Sale) Systems, etc. Usually, these source

systems keep the data for a limited time and afterwards the data is being archived. Studies in business intelligence field revealed that exploiting transactional data leads to a useful business analyze. The transactional data becomes a data source for the data warehouse, another important component of the business intelligence system. Depending on the organizations, business intelligence needs and the systems purpose, external data, such as other organizations data or statistic data, can be selected as reliable data source. At the end of the analyzing stage, analysts must have a general point of view over the organization targets.

b) Data warehouses

The analyzing and implementation phase of the data warehouses, it's for sure the most important phase in developing the business intelligence system and it's closely correlated with other phases. In this stage, it will established how the ETL process is going to be realized, how it will be implemented the OLAP cube or in what technology will the reports be made in. [5]

So, a data warehouse is still a database, mostly relational, which is implemented differently than the standard databases and contains historical data of a certain interest. Ralph Kimball, known for his interest in business intelligence, defined the data warehouse as "a copy of transactional systems, especially structured for querying and analysis". [6]

Depending on the coverage, the warehouses can be divided in Enterprise Warehouse, Data Marts and Virtual Warehouses. The Enterprise Warehouse collects all the information, which regards the whole organization. Usually, this provides very big volumes of data, containing detailed data, and also aggregates data, and as dimensions, most of the times it reaches tens of terabytes. Data concentrations or data marts contain only a subset of the data volume from the organization, specific to a certain group of users, being limited to specific subjects. The data contained by those data concentrations, are usually aggregate data. The virtual warehouse is "a set of visions of the operational databases" [7], being easy enough to implement, but needing supplementary capacities on the data servers.

A data warehouse usually contains aggregate data, detailed data and metadata.

Aggregate data, even they determine a rise of data redundancies, are necessary in a data warehouse, because they improve the systems average response time. These assume a point of prior processing, so that they will be ready for the management's needs: they can be consolidated, totalized, summarized.

Detailed data is the relative recent data, delivered to the user, usually to execution level.

c) Extracting, Transforming and Loading Data

The ETL (Extract, Load, Transform) is one of the main components of a business intelligence system, on which it depends in highly measure, the data accuracy the organization will analyze. [8]

The data extraction will be a very laborious process and many times can be a challenge for developers, because this stage depends on the good functioning of the business intelligence system. Data which will be extracted in this stage is the one loaded in the data warehouse and on which the transformations and reports will be made in the next stages. For this reason, it is very important that the data is correctly extracted from the source files. Most of the times, the data is constituted from data files (flat files), exported from relational databases, but there are situations in which external data has to be loaded in the data warehouse. The big challenge is the transformation process of all the data resources in one single type,

accepted by the data warehouses. The ETL process, during its extraction phase, will take care of this problem. It will "inspect" the data files, will check if they are in a certain format, will load the data in an initial database and will reject all the incorrect data. A first data filter that is to be loaded is made in this first stage.

The data transformation stage is made from a set of rules and functions that apply to the extracted data in the first step. The purpose of this process is to prepare the data for the loading stage. Depending on the business intelligence system requirements, this stage can imply very complex transformations of the extracted data, or, on the contrary, they can be very little refined; all depend, of course, on what the systems has to fulfill.

The transformation stage is, like the data extraction stage, a very laborious and challenging process, being one of the key stages in implementing a business intelligence system. In this stage the data transforms in relevant information for the organizations management.

The data loading stage represents the last of the three ETL subsystem stages and it consists in loading data in the target tables. Of course that during this stage, the situation can be different from an organization to another, depending on the business requirements. Some companies can request that this loading to be made at certain time moments; data can be loaded daily, monthly, annualy etc. For some data warehouses there can be an initial loading, for others they can only make updates or complementary data insertions.

d) Multidimensional Data Analyze

OLAP (On Line Analytical Processing) is a well known technology in business intelligence. This technology has its roots in complex analyzing and processing methods, they look as a ensemble of dimensions, hierarchical and interrelated measures [9]. The OLAP technology offers, first of all, system performances, meaning that data is aggregated and as a multidimensional cube.

At present there are different types of OLAP analyzing, and the software developers come with multiple technologies for multidimensional analyze. OLAP analyzing instruments allow the elaboration of complex reports, but for viewing the reports special designed instruments can be used.

Conclusions

Business intelligence is, at this time, one of IT fields with continuous improvements. Starting with basic theoretical notions and ending with the new technologies that are developing this way, business intelligence concepts are in the sight of all IT people.

As we mentioned before, business intelligence appeared as an answer to the economics' environment needs. Big organizations and multinational companies have already implemented a solution for business intelligence. Though implementing this kind of system is extremely laborious and expensive, the benefits proved to be many more. As it was proven in the paper, integrating storing and exploiting an organization's data can bring important advantages.

The main advantage in using business intelligence is the ability to transform data into information. This allows companies to develop an efficient mechanism of making decisions, in other words to make better and faster decisions. The benefits can be important for the company's management in making strategic decisions, but it can also help department leaders, analysts or any other member of a team faced with having to take decisions.

Analyzing intelligent data has always been important because through this analyze intelligence is being generated. Business intelligence is like an endless story, applicable in fields like audit, risk management, diplomacy or lobby

activities. And new fields are just shaping up. This would be the case of merge between business intelligence and artificial intelligence, merger that would lead to a new concept, artificial business intelligence (ABI).

In these conditions we must admit that these technologies are just at the beginning of a long journey, in a world where the key of success stands within the capacity of making better decisions in a shorter time then the competition. Besides, a company's life depends more and more on such decisions, which will make impossible not to admit the benefit brought by business intelligence.

References

[1] John C. Hancock, Roger Toren – Practical Business Intelligence with SQL Server 2005", Addison Wesley Professional, 2006

[2] Gartner – Essential Components and Success Factors of Business Intelligence and Performance Management. Cannes, France: Gartner Symposium IT Expo 2006

[3] Wikipedia.org – http://en.wikipedia.org/wiki/Business_Intelligence

[4] Architecture for Enterprise Business Intelligence – Microstrategy http://www2.microstrategy.com/download/files/whitepapers/open/Architecture-for-Enterprise-BI.pdf

[5] Data Warehousing - http://www.1keydata.com/datawarehousing/datawarehouse.html

[6] Ralph Kimball – The Data Warehouse Lifecycle Toolkit (2nd ed.), 2008

[7] M. Velicanu, I. Lungu, I. Botha, A. Bara, A. Velicanu, E. Rednic – "Sisteme de baze de date evaluate", Bucuresti, 2009

[8] Wikipedia.org - http://en.wikipedia.org/wiki/Extract,_transform,_load

[9] On-Line Analytical Processing – www.olap.com

Bogdan NEDELCU graduated Computer Science at Politehnica University of Bucharest in 2011. In 2013, he graduated the master program "Engineering and Business Management Systems" at Politehnica University of Bucharest. At present he is studying for the doctor's degree at the Academy of Economic Studies from Bucharest.

Data Mining Solutions for the Business Environment

Ruxandra PETRE
University of Economic Studies, Bucharest, Romania
ruxandra_stefania.petre@yahoo.com

Over the past years, data mining became a matter of considerable importance due to the large amounts of data available in the applications belonging to various domains. Data mining, a dynamic and fast-expanding field, that applies advanced data analysis techniques, from statistics, machine learning, database systems or artificial intelligence, in order to discover relevant patterns, trends and relations contained within the data, information impossible to observe using other techniques.
The paper focuses on presenting the applications of data mining in the business environment. It contains a general overview of data mining, providing a definition of the concept, enumerating six primary data mining techniques and mentioning the main fields for which data mining can be applied. The paper also presents the main business areas which can benefit from the use of data mining tools, along with their use cases: retail, banking and insurance. Also the main commercially available data mining tools and their key features are presented within the paper.
Besides the analysis of data mining and the business areas that can successfully apply it, the paper presents the main features of a data mining solution that can be applied for the business environment and the architecture, with its main components, for the solution, that would help improve customer experiences and decision-making.
Keywords: *Data mining, Business, Architecture, Data warehouse*

1 Introduction

Nowadays, companies collect huge volumes of data on a daily basis. Analyzing this data and discovering the meaningful information contained by it became an essential need for businesses.
As the business environment develops and changes constantly, facing every day new challenges, the companies try to strengthen their market position and achieve competitive advantage by using new and innovative solutions, like data mining.
Data mining solutions implement advanced data analysis techniques used by companies for discovering unexpected patterns extracted from vast amounts of data, patterns that offer relevant knowledge for predicting future outcomes.

2. General overview of data mining

The availability and affluence of data belonging to various domains make data analysis a matter of significant importance and necessity today. Data mining – the analysis step within the KDD (Knowledge Discovery in Databases) process – uses a diversity of advanced data analysis methods to explore the data and discover useful patterns and trends.
Data mining consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data. [1]
With the imminent growth of the amounts of data in every application, using data mining methods for automatically identifying valid and meaningful patterns in order to produce useful information and knowledge became a requirement for various fields including business, education or science and engineering, fields for which data mining can fulfill the following purposes:

- *Business* – data mining can be applied in retail, banking or insurances, for activities like customer segmentation

and retention, market basket analysis or fraud detection;
- *Education* – data mining can be applied for grouping students, predicting student performance, planning and scheduling courses or understanding student behavior;
- *Science and engineering* – data mining can be used for domains like bioinformatics, astronomy, medicine, genetics, electrical power, telecommunications or climate data.

Data mining can be defined as a process of exploring and analysis for large amounts of data with a specific target on discovering significantly important patterns and rules. Data mining helps finding knowledge from raw, unprocessed data. Using data mining techniques allows extracting knowledge from the data mart, data warehouse and, in particular cases, even from operational databases. [2]

The data mining methods, used for extracting hidden patterns from the data, are classified into the following two categories: description methods and prediction methods. Description methods are oriented to data interpretation, which focuses on understanding (by visualization for example) the way the underlying data relates to its parts. Prediction-oriented methods aim to automatically build a behavioral model, which obtains new and unseen samples and is able to predict values of one or more variables related to the sample. [3]

Data mining analyzes the data by applying a wide variety of techniques, developed for the efficient handling of large volumes of data. The six primary data mining techniques are presented below in figure 1:

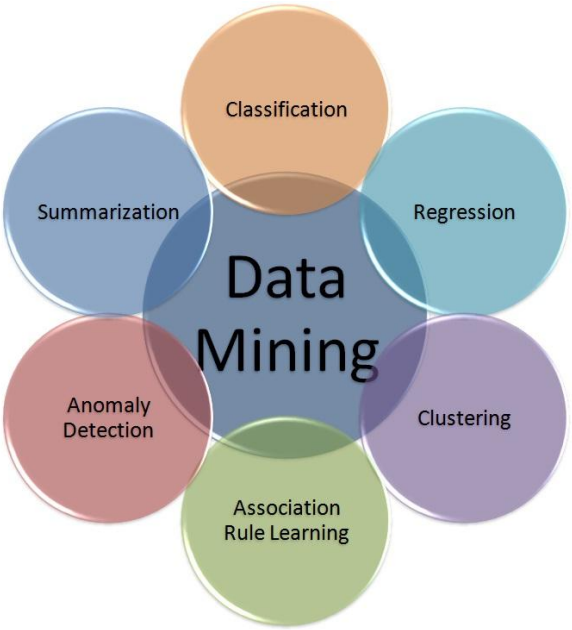


Fig. 1 Data mining techniques

The main data mining techniques are organized into the following categories: [1]
- *Classification*: consists of a function that maps (classifies) a data item into one of several predefined classes;
- *Regression*: involves a function that maps a data item to a real-valued prediction variable;
- *Clustering*: is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data;
- *Association rule learning (Dependency modeling)*: consists of finding a model that describes significant dependencies between variables;

- *Anomaly detection (Change and deviation detection)*: focuses on discovering the most significant changes in the data from previously measured or normative values;
- *Summarization*: involves methods for finding a compact description for a subset of data.

Data mining has evolved in the past two decades, becoming a fundamental discovery process. It has incorporated techniques from many other fields, including statistics, machine learning and database systems.

The diversity of data and the multitude of data mining techniques provide various applications for data mining, which have improved many domains of human life.

3. Data mining applications for business

Data mining is defined as a business process for exploring large amounts of data to discover meaningful patterns and rules. [4]

Companies can apply data mining in order to improve their business and gain advantages over the competitors.

The most important business areas that successfully apply data mining, presented in Fig. 2 below, are:

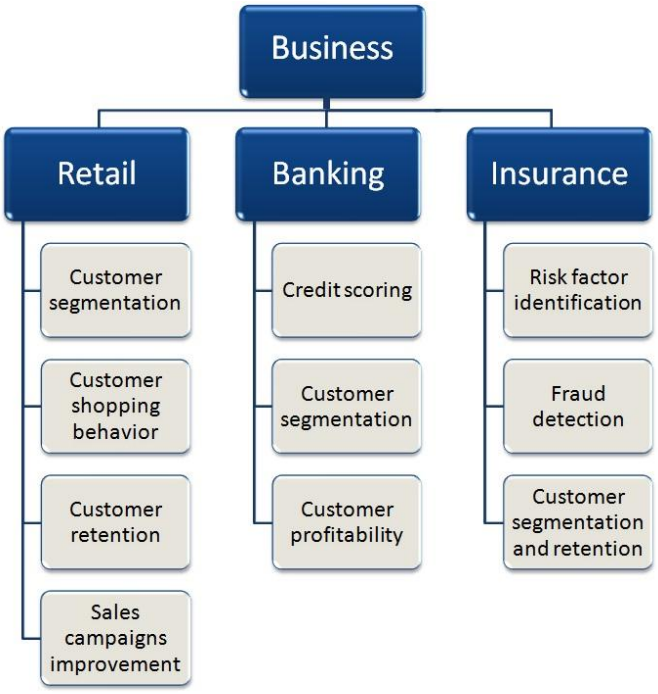


Fig. 2 Business areas that successfully apply data mining

1. *Retail*

Retail data mining can help identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business. [5]

Data mining techniques have many applications in the retail industry, including the following:

- *Customer segmentation*: identify customer groups and associate each customer to the proper group;
- *Establish customer shopping behavior*: identify customer buying patterns and determine what products the customer is likely to buy next;
- *Customer retention*: identify customer shopping patterns and adjust the

product portfolio, the pricing and the promotions offered;
- *Analyze sales campaigns*: predict the effectiveness of a sales campaign based on the certain factors, like the discounts offered or the advertisements used.

Retail industry offers a wide area of applications for data mining due to the large amounts of data available for companies.

2. *Banking*

There are various areas in which data mining can be used in financial sectors like customer segmentation and profitability, credit analysis, predicting payment default, marketing, fraudulent transactions, ranking investments, optimizing stock portfolios, cash management and forecasting operations, high risk loan applicants, most profitable Credit Card Customers and Cross Selling. [6]

The main examples of applications of the data mining techniques in the banking industry are the following:
- *Credit scoring*: distinguish the factors, like customer payment history, that can have a higher or lower influence over loan payment;
- *Customer segmentation*: establish customer groups and include each new customer in the right group;
- *Customer retention*: identify customer shopping patterns and adjust the product portfolio, the pricing and the promotions offered;
- *Predict customer profitability*: identify patterns based on various factors, like products used by a customer, in order to predict the profitability of the customer.

The information systems for the banking industry contain large amounts of operational and historical data, being a fitted application area for data mining.

3. *Insurance*.

Data mining can help insurance firms in business practices such as: acquiring new customers, retaining existing customers, performing sophisticated classification or correlation between policy designing and policy selection. [7]

In insurance the data mining techniques have the following applications:
- *Risk factor identification*: analyze the factors, like customer claims history or behavior patterns, that can have a stronger or weaker influence over the insured's level of risk;
- *Fraud detection*: establish patterns of fraud and analyze the factors that indicate a high probability of fraud for a claim;
- *Customer segmentation and retention*: establish customer groups and include each new customer to the appropriate group and identify discounts and packages that would increase customer loyalty.

Data mining techniques have many applications in the insurance business and can improve it by analyzing the large amounts of data available for companies.

4. Data mining tools used in the business environment

Data mining tools commercially available implement various data mining techniques for performing advanced data analysis on large volumes of data. The main data mining products, presented in Table 1 below, along with their key features, are: IBM SPSS Modeler, developed by IBM, the data mining tools included by Microsoft SQL Server Analysis Services, Oracle Data Mining, embedded within the Oracle database, SAS Enterprise Miner, produced by SAS, and STATISTICA Data Miner, developed by StatSoft.

Table 1 Main commercially available data mining tools and their key features

Data mining tool	Key features
IBM SPSS Modeler	• Data mining and text analytics software application used for building predictive models • Intuitive graphic user interface that allows users to import, manage and analyze their data • Data mining techniques included are: clustering (K-means, Support Vector Machine), classification (Bayesian networks, regression, neural networks, decision trees), association rules (Apriori), anomaly detection • Application for which it can be used: forecasting sales, customer relationship management, risk management or fraud detection
Microsoft SQL Server Analysis Services	• OLAP, data mining and reporting tool in Microsoft SQL Server • Used to create, manage, and explore data mining models, and then create predictions by using those models • Data mining algorithm types included are: classification, regression, clustering, association algorithms, sequence analysis • Tasks for which it can be used: customer segmentation, forecasting sales, market basket analysis, identifying customer shopping behavior
Oracle Data Mining	• Embeds data mining techniques within the Oracle database • Provides means for building, testing, validating, managing and deploying data mining models inside the database environment • Supports the following data mining functions: classification, regression, attribute importance, anomaly detection, clustering, association models and feature extraction • Applications for which it can be used: customer segmentation, recommend next likely product, credit scoring, customer profitability or fraud detection
SAS Enterprise Miner	• Software application that provides data mining algorithms for creating predictive and descriptive models • Comprises an easy to use graphical user interface that helps with data preparation, summarization and exploration, as well as advanced predictive and descriptive modeling • Data mining techniques applied include: classification (decision trees, neural networks), clustering, regression, association rules • Tasks for which it can be applied: detect fraud, anticipate resource demands, increase acquisitions and curb customer attrition.
STATISTICA Data Miner	• Statistics and analytics software package that provides data analysis, data management, statistics, data mining and data

	visualization functions • Provides effective data pre-processing, cleaning, and filtering tools, along with tools for producing prediction models in various formats • Data mining methods available include: clustering, classification, regression, association and sequence analysis • Applications for which it can be used: customer segmentation, customer retention, credit scoring, market basket analysis or price optimization

Several researchers and organizations have conducted reviews and surveys of data mining tools. These examine and provide an overview of the behaviors, preferences and views of data mining, data science and analytic professionals.

One of these reports is Annual Rexer Analytics Data Miner Surveys, published by Rexer Analytics.

The selection of the primary commercial data mining tools in 2013, according to this report, is presented in Fig. 3 below:

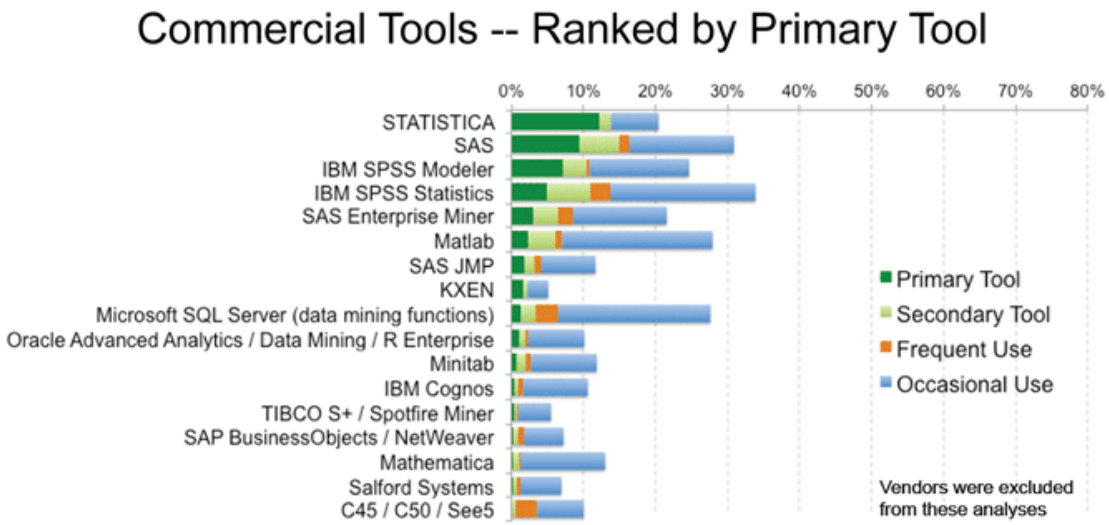


Fig. 3 Primary commercial data mining tools in 2013 [8]

The analysis of the above figure shows for the commercial data mining tools, which tools were considered by the respondents as primary tool, secondary tool, frequently or occasionally used tool. According to this survey, in 2013, the primary data mining tools used were STATISTICA Data Miner, IBM SPSS Modeler and SAS Enterprise Miner.

Each data mining tool analyzed has different features and can be used for various requirements.

5. Data mining solution for the business environment

Business is well-fit domain for applying data mining as it provides large volumes of data.

The main features of a data mining solution for the business environment, presented in Fig. 4 below, are:

Fig. 4 Main features of a data mining solution for business

- *Selecting the data*: identify the data sets used for a specific analysis and improve the initially selected data sets if required;
- *Preparing the data*: transform and clean the data so it is in the appropriate format for applying data mining techniques;
- *Choosing the data mining technique*: select the algorithm associated to a data mining technique that is suitable for the required analysis;
- *Configuring the settings for the data mining technique*: configure for the selected algorithm the necessary parameters;
- *Executing the data mining process*: execute the configured data mining process;
- *Viewing the results of the data mining process*: view the results generated by the execution of the data mining process.

The architecture of the proposed data mining solution, applied to the business environment, is illustrated in Fig. 5 below:

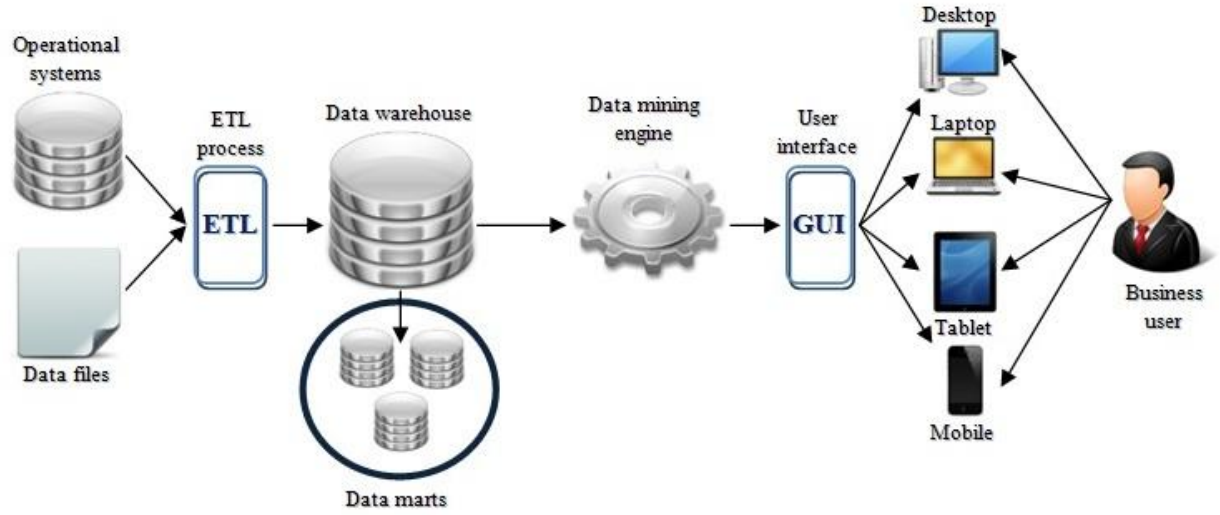


Fig. 5 Architecture of data mining solution for business

The data is extracted from the data sources, both operational systems and data files, and loaded through the ETL (Extract, Transform and Load) process to the data warehouse. The data warehouse can contain raw data – the data in a detailed format, as it has been extracted, summary data – data that has been aggregated and transformed – and metadata, data that provides information about the raw and summary data belonging to the data warehouse. The data belonging to the data warehouse can be organized in data marts.

The solution provides a data mining engine that may be used for obtaining advanced analysis. The solution has a graphical user interface that provides access to the main features of the solution, as presented in the article: select and prepare the data, choose data mining technique, configure the settings and execute the data mining process, view the results obtained.

The business user may access the GUI of the solution from various devices, like the desktop, laptop, tablet or mobile.

Using the functionalities described above the solution allows accessing and analyzing business related information in order to obtain valuable knowledge concerning the business.

Conclusions

Our current society needs data mining for improving many domains of human life. Business areas like retail, banking and insurance can use data mining methods to improve customer experiences, make optimal decisions, strengthen their market position and achieve competitive advantage.

There are various commercially available data mining tools to provide support for fulfilling these requirements.

The architecture proposed for the data mining solution for the business environment would improve the efficiency of a company, by providing valuable decision-making knowledge to minimize operating costs and gain competitive advantage.

References

[1] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, *From Data Mining to Knowledge Discovery in Databases*, AI Magazine, Vol. 17, Issue 3, 1996, ISSN 0738-4602, pp. 37-54.

[2] Ion Lungu and Adela Bâra – "Improving Decision Support Systems with Data Mining Techniques", "Advances in Data Mining Knowledge Discovery and Applications" – chapter 18, InTech Publisher, Croatia, 2012, ISBN 978-953-51-0748-4, pp. 397-418.

[3] Oded Maimon and Lior Rokach, *Data Mining and Knowledge Discovery Handbook. Second Edition*, Springer Publishing Company, USA, 2010.

[4] Gordon S. Linoff and Michael J. A. Berry, *Data Mining Techniques: for Marketing, Sales and Customer Relationship Management. Third Edition*, Wiley Publishing, USA, 2011.

[5] Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques. Third Edition*, Morgan Kaufmann Publishing, USA, 2011.

[6] Vikas Jayasree and Rethnamoney Vijayalakshmi Siva Balan, *A Review on Data Mining in Banking Sector*, American Journal of Applied Sciences, Vol. 10, Issue 10, 2013, ISSN 1554-3641, pp. 1160-1165.

[7] A. B. Devale and Dr. R. V. Kulkarni, *Applications of data mining techniques in life insurance*, International Journal of Data Mining & Knowledge Management Process, Vol.2, Issue 4, July 2012, ISSN 2230-9608, pp. 31-40.

[8] Rexer Analytics, *Annual Rexer Analytics Data Miner Survey*, Commercial Tools – Ranked by Primary Tool, 2013, Available: http://www.statsoft.com/Company/About-Us/Reviews/2013-Published-Reviews.

Ruxandra PETRE graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2010. In 2012 she graduated the Business Support Databases Master program. Currently, she is a PhD candidate, coordinated by Professor Ion LUNGU in the field of Economic Informatics at the Bucharest University of Economic Studies. The title of her PhD thesis is "Information Solutions for Data Analysis". Her scientific fields of interest include: Databases, Data Warehouses, Business Intelligence, Decision Support Systems and Data Mining.

Big Data and Specific Analysis Methods for Insurance Fraud Detection

Ana-Ramona BOLOGA, Razvan BOLOGA, Alexandra FLOREA
University of Economic Studies, Bucharest, Romania
ramona.bologa@ie.ase.ro, razvanbologa@ase.ro, alexandra.florea@ie.ase.ro

Analytics is the future of big data because only transforming data into information gives them value and can turn data in business in competitive advantage. Large data volumes, their variety and the increasing speed their growth, stretch the boundaries of traditional data warehouses and ETL tools. This paper investigates the benefits of Big Data technology and main methods of analysis that can be applied to the particular case of fraud detection in public health insurance system in Romania.
Keywords: *Big Data, Social Networks, Data Mining, Fraud Detection*

1 Introduction

Health budgets are a common target of fraudulent practices. Due to the complicated nature of medical processes, frauds have always found a favorable environment in the health insurance system.

Since fraud is on, the increase holistic fraud prevention is required. According to the well know market research organization Gartner [9]: "Security and fraud risk exposure is increasing as organizations are threatened at multiple points of vulnerability. Companies are re-evaluating how they tackle security since a fragmented approach is consistently leaving organizations at greater risk of attack. A more holistic approach to security ensures all layers of protection function together".

Electronic health cards with smart chips have been implemented in order to fight fraud in health insurance. The use of eHealth cards has been generating a huge amount of data that needs to be processed. The conventional database technologies are not suitable for performing this type of analysis due to their inner limitations. The new big data technologies are yet to be understood and the benefits they provide in fighting fraud need to be investigated.

By performing big data analysis, common repetitive errors that are "hidden" inside huge repositories of data can be identified and corrected. Such errors would go undetected in the absence of big data technologies because the human brain is not capable to correlate the huge quantities of data available in the medical sector.

In order to prevent insurance fraud, big data analytics should use the following technologies: business rules, anomaly detection, text mining, database searches and social network analysis. These technologies will be approached during the following sections.

2. Big data technology - advantages and challenges

The subject of big data is of major interest to the scientific community as the size of the databases has been growing beyond the limits of current technologies. As indicated in the bibliography section, there is consistent research of big data technologies with applications into the health sector.

There are currently many research initiatives for developing big data technologies. Some of these initiatives are funded by private companies such as IBM, ORACLE, SAS and Microsoft. Other initiatives are funded by public bodies and/or the open source community. A notable technology open source is Hadoop which is often integrated with commercial technologies.

However, the applications of such technologies are still to be developed as they

are highly dependent on each sector of activity and on each geographical area.

Although massive data amounts were produced during the last two years, the term "big data" was present in the research literature starting with 1970s, but it has seen an explosion of publications since 2008 [3].

Wikipedia defines big data as "a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications" [12].

The industry standard definition of Big Data projects it along four dimensions: volume, velocity, variety and veracity [17].

Volume dimension refers to the huge data volume produced or manipulated by a company that must be further manipulated in order to get useful information.

Velocity dimension refers to speed of data processing, as some activities need real-time responses. Distributed and parallel processing algorithms become very important for this reason. During the fraud detection process it is very important to analyze day-by-day big data flows, millions of detailed record that must be scrutinized to get a behavior pattern identification.

Variety dimension refers to various types of data that are manipulated by a company, both structured and unstructured (text, audio, video, click streams, log files, sensor data etc.).

Veracity dimension refers to the information level of trust granted by the business decision factors.

More than that, when working with big data, the meaning of each event can be interpreted only in relationship with preceding events. So, we heave streams of data that must be analyzed all together, like a sequence, so the traditional analytic methods work poorly on these cases. Traditional analytic tools approach data at entity level, as each entity provides useful information. The shift to detailed stream data changes the needs and requires for complex ETL tools.

First used by Internet giants like Yahoo, Ebay or Facebook, Apache Hadoop is the most popular big data platform. Hadoop is an open source platform for processing big data that uses distributed processing across clusters of servers. It has become the "de facto" standard for storing, processing and analyzing huge amounts of data.

Hadoop is a Java based framework and uses simple parallel programming models using clusters of inexpensive servers that locally store and process huge volumes of data. The result is a fundamental decrease of data storage cost. The analysts are free to write code in almost any contemporary language using the streaming APIs available in Hadoop.

The platform offers a high level of scalability as processing requirements are distributed on thousands of machines and its software is designed to detect and solve failures at application level. This way, its clusters are very resilient.

Core Hadoop has two main systems:

- ***Hadoop Distributed File System (HDFS)***: self-healing high-bandwidth clustered storage.
- ***MapReduce***: distributed fault-tolerant resource management and scheduling coupled with a scalable data programming abstraction.

In the beginning (2008), Hadoop had significantly less capabilities then relational databases and had limited supporting tools. But now, it has more robust SQL capabilities and access to all SQL-based applications. Cloudera was the first that introduced commercial support for Hadoop in 2008, followed by MapR and Hortonworks. IBM and EMC have each its own Hadoop distribution. Microsoft and Teradata offer complementary software Hortonworks' platform. Oracle resells and supports Cloudera, while HP, SAP work with multiple Hadoop software providers [16].

Classic business intelligence tools use

relational databases for storage and query execution. In order to use the traditional analysis methods and techniques, there have been many efforts to develop SQL-like languages for big data access, query and manipulation: BigSQL, HiveQL, CassandraQL, JAQL, Sparql, Shark etc, each of them associated with a specific big data platform.

Many users concluded that no type of big data is optimal for all their requirements. Today there are many implementations of hybrid big data architecture, which combine two or more technologies in specialized roles (see Fig.1). For example, combining Hadoop for unstructured data staging with in-memory business intelligence tools for query acceleration, with stream computing for continuous data provision and with massively parallel processing RDBMS for data warehousing and data management [18].

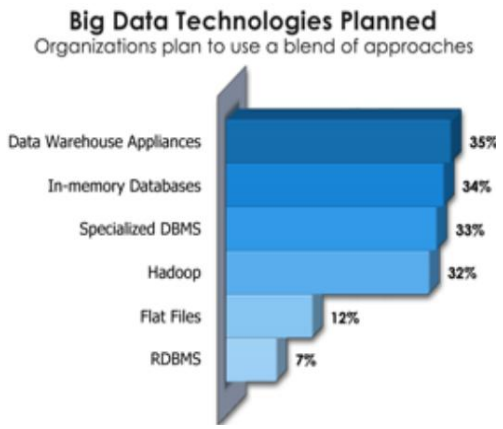


Fig. 1. Hybrid technology approach [21]

Big Data Connectors are used to combine RDBMS with Hadoop for deeper analysis, using data mining or statistical analysis.

More than that, in order to enable more flexible topologies of Hadoop and non-Hadoop solutions, a standard query virtualization layer can be developed to support a transparent SQL access to any platform [18].

A scan of the list of projects reported as using big data techniques by PoweredBy Hadoop [20], suggests that the big data approach is best suited for business problems that meet one or more of the following criteria [5]:

- Data-restricted throttling
- Computation-restricted throttling
- Large data volumes
- Significant data variety
- Benefits from data parallelization.

The use of big data for fraud prevention has a huge potential as the data generated by the current transactional systems is enormous and current database technologies are unable to process it. All the five criteria listed before are respected, which make fraud prevention a perfect suited application for big data analytics.

The next section will address traditional methods of analysis for the detection of fraud and how these can be exploited in the context of the big

3. Fraud in the health insurance system in Romania

The main issue in fraud detection is the fact that the collections of data are impossible to be processed by a human brain. For instance, a controller could observe that in a short period of time all the inhabitants located on a single street did a set of expensive laboratory tests, say for hormonal disorders, which a medical lab is charging to the health insurance system. This is clearly a fraud as it highly unlikely that a very limited number of persons located on a single street go, in a short period of time, to take such rare and expensive tests.

It is easy to suspect that the doctor who signed the test results probably did not act on his own and that such frauds most probably happened before. The controller might want to analyze not only the activity of the doctor that signed the test results, but also the activity of the persons from his/her social network (colleagues, managers, previous managers and others).

A tool that allows such analysis for the health insurance system does not exist on the current market. During discussions with global leading companies, the Romanian Health Insurance Agency discovered that

such tools exist for the banking industry but nobody adapted them to the health sector.

National Health Insurance System in Romania has been continuously restructured during the last 20 years and the following paragraphs summarize the main legislation relating to these changes, according to the site of the National House of Health Insurance [19].

The Law Social Health Insurance - Law no. 145/1997 was adopted in June 1997. This followed the type Bismarck insurance model with compulsory health insurance based on the principle of solidarity and operating under a decentralized system. It came into force on January 1, 1999. In consequence, from 1 January 1999, Insurance Houses have functioned as autonomous public institutions, led by representatives of the insured and employers through the boards, as well as the National House of Health Insurance (19).

O.U.G.no.150/20.11.2002 on "Organization and functioning of health insurance" repealed law no. 145/1997. This allowed conceptual and structural changes of the health insurance system as a unified system of financing care and promoting health. According to this normative act, social health insurance system in Romania has three major components:

- Insured person;
- Health care providers (doctors, hospitals, pharmacies);
- Health insurance houses (tertiary payer).

Law no. 95/2006 has produced a new health system reform in Romania, imposing more flexibility and dynamism, giving clear responsibilities to ensure both logistics for the coordinated functioning of the health insurance system (by collecting and efficient use of funds), and appropriate means for representing, informing and supporting the interests of the insured.

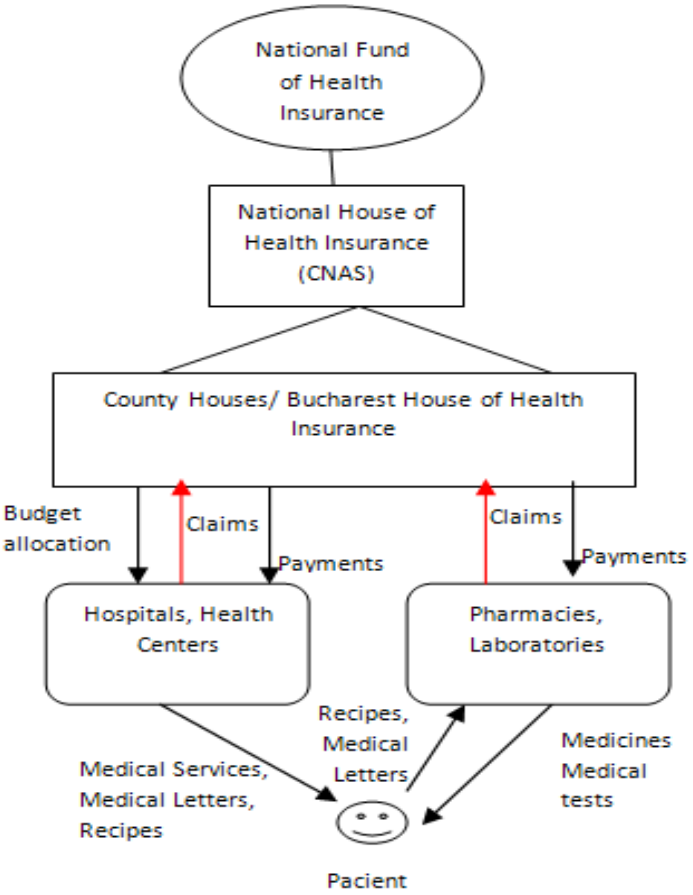


Fig. 2. Main entities involved Health Insurance System

Basic principles for the organization and functioning of the National House of Health Insurance are:

- Free choice of health insurance house;
- Solidarity and subsidiarity in the collection and use of funds;
- Free choice of family doctor, physician and health unit;
- Mandatory participation to health insurance contribution payment for the formation of National Fund of Health Insurance;
- Participation of insured persons, employers and government to the National Fund of Health Insurance management;
- Provide a package of basic health services, equitably and without discrimination of any insured;
- Transparency health insurance system.

From a financial perspective, the Budget of the National Health Insurance Fund is approved by the Annual State Budget Law and takes into consideration:

- Current requirement of medical activity;
- Amounts representing arrears recorded in hospitals and open circuit pharmacies;
- Financial resources in each county.

In Figure 2, flows which are recording most cases of fraud are marked with red color. The main types of fraud that can be identified in the Romanian insurance health system are mostly similar to other European health systems [5]:

- Unusual high number of invoices for a particular insured person in a short time (3-4 days);
- Use of false identities for claiming false hospitalization, false prescriptions or other false health care services;
- Claiming medical invoices having dates outside the insurance period;
- Excessive number of medical claims in a certain period;
- An excessive number of manual invoices requests whose values are usually lower than the limit of inspection;
- Claims having payable amounts higher than the billed amounts that insurance house will pay.

4. Analysis methods for detecting fraud in health insurance

Insurance fraud can be defined as "knowingly making a fictitious claim, inflating a claim or adding extra items to a claim, or being in any way dishonest with the intention of gaining more than legitimate entitlement" [2].

The current health issuance fraud is about 5% of the health budgets. In Romania alone this amount represents around 250-300 million Euros/year [15]. Most of this fraud would be detectable by clever data analytics. The area of fraud prevention has been traditionally correlated with data mining and text mining. Even before the "big data" phenomena started in 2008, text mining and data mining were used as instruments of fraud detection. However, the limited technological capabilities of the pre-big data technologies made it very difficult for researchers to run fraud detection algorithms on large amounts of data.

Frauds in Health Insurance system can be specific to each country, usually based on gaps or weaknesses of legislation. Models are constantly changing fraud, malicious individuals seeking ever new ways to circumvent the law. Consequently, methods for identifying and preventing fraud must always be adjusted and ready to rediscover the fraudulent actions.

In general we can identify two types of fraud [13]:

1. ***Opportunistic fraud,*** when a person takes advantage of the deliberate padding or inflating of a legitimate insurance claim. This type of fraud is very common, but the incident is related to a reduced amount.
2. ***Professional fraud***, usually done by organized groups of people who may have multiple, false identities. They know very well how to organize the

system and often work together with people within the system. The incidence of these events is lower, but the amount related to an incident is much higher.

The data used for analysis are taken from the database of the National House for Health Insurance and contains all necessary information on the partners involved in events claim payments for medical services.

Specific attributes are used to detect frauds that are usually the same. Thus, in the field of health insurance it can be taken into account: patient demographics (age, gender), details of the medical services provided to the patient, and details of the claim. [11]

The complex nature of the data used in fraud detection has been well described by [1]:

- Volume of both fraud and legal classes will fluctuate independently of each other; therefore class distributions (proportion of illegitimate examples to legitimate examples) will change over time.
- Multiple styles of fraud can happen at around the same time. Each style can have a regular, occasional, seasonal, or once-off temporal characteristic;
- Legal characteristics/behavior can change over time.
- Within the near future after uncovering the current modus operandi of professional fraudsters, these same fraudsters will continually supply new or modified styles of fraud until the detection systems start generating false negatives again.

Techniques used for fraud detection fall into two primary classes: statistical techniques and artificial intelligence.[7] Examples of ***statistical data analysis techniques*** are:

a. Data preprocessing techniques for detection, validation, error correction, and filling up of missing or incorrect data.
b. Calculation of various statistical parameters such as averages, quantiles, performance metrics, probability distributions, and so on. For example, the averages may include average length of call, average number of calls per month and average delays in bill payment.
c. Models and probability distributions of various business activities either in terms of various parameters or probability distributions.
d. Computing user profiles.
e. Time-series analysis of time-dependent data.
f. Clustering and classification to find patterns and associations among groups of data.
g. Matching algorithms to detect anomalies in the behavior of transactions or users as compared to previously known models and profiles. Techniques are also needed to eliminate false alarms, estimate risks, and predict future of current transactions or users.

Fraud management is a knowledge-intensive activity. The main ***AI techniques*** used for fraud management include [AI]:

a. Data mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud.
b. Expert systems to encode expertise for detecting fraud in the form of rules.
c. Pattern recognition to detect approximate classes, clusters, or patterns of suspicious behavior either automatically (unsupervised) or to match given inputs.
d. Machine learning techniques to automatically identify characteristics of fraud.
e. Neural networks that can learn suspicious patterns from samples and used later to detect them.

Figure 3 shows the analysis methods depending on the types of fraud and the types of frauders [SAS1].

Data mining techniques can be used for fraud detection for large sets of data from health insurance system. These techniques detect behavior patterns in large data sets, so based on several cases considered fraudulent can calculate the probability that each record be fraudulent.

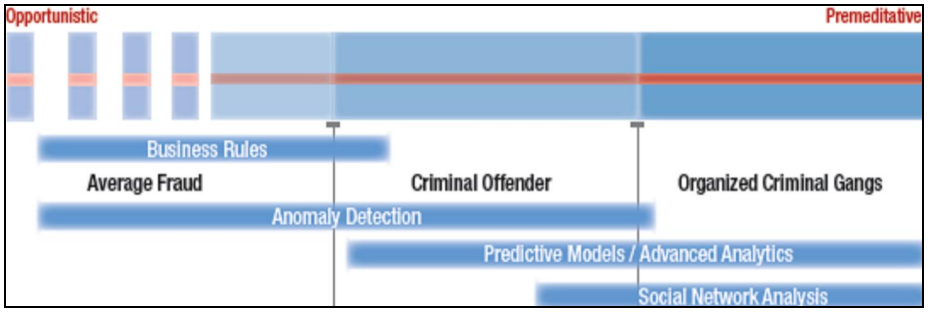


Fig. 3. Techniques for fraud detection ([SAS1])

This data must be available, relevant, adequate and clean. There are two main criticisms of data-mining fraud detection tools: the dearth of publicly available data for analysis and the lack of published well-known methods and techniques that are specifically efficient for this field [8].

One of the most commonly used techniques for detecting fraud is **anomaly detection.**

Anomaly detection algorithms are very simple to set and functions automatically. Some key performance indicators are for an event chosen and then thresholds are set. If a threshold is exceeded, then the event is signaled for further investigation. The effectiveness of this method is influenced by the choice of indicators to be monitored, of the analysis period, and of the threshold value settings.

Business Rules

If fraud patterns are known, one can resort to checking every transaction by applying business rules. Based on an aggregate score or exceeding a set threshold, a transaction can and marked as suspicious, and then carefully investigated. Figure 4 presents an example of a business rule used for validation in the claim processing application:

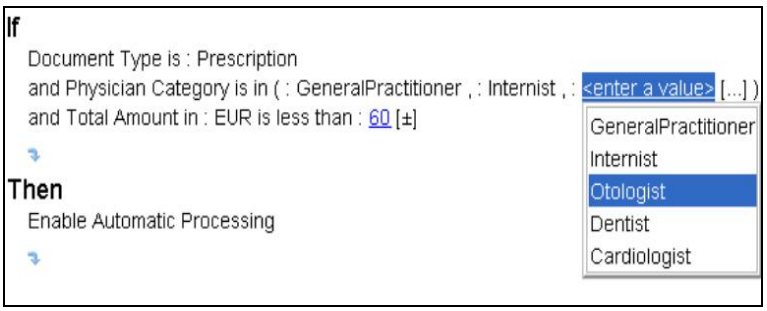


Fig. 4. Example of business rule [4]

This technique is very simple to apply, once the system was originally set. Its weaknesses are two: setting initial parameters can lead to many false alarms that require further investigation, and the system is flexible to adapt to new methods to defraud the system, new business rules. You can add new business rules only if they meet the new method of fraud.

Database searching

For records detected as suspicious further investigation. One approach is the use of the database searching services, which can give investigators a large amount of information from multiple sources. Was the suspicious person involved in illegal

activities? Had he attempted fraud in other areas in the past? Information can be obtained by searching the data to other companies that can help solve the case.

Predictive modeling

Predictive modeling is very successful in detecting fraud. By applying data mining tools, fraud propensity scores can be calculated. Then, using predictive models, they can automatically tell the probability that data is fraudulent and it must be subjected to detailed analysis.

To preserve accuracy, models must be constantly updated to include new types of illegal events.

Text mining

Text mining is a very useful technique as almost 80% of data generated by insurance claiming process have an unstructured form. This technique is very efficient on big data volumes. Meaningful data are extracted and then analyzed by text mining algorithms to reveal abnormal or suspicious behavior of the insured.

Social network analysis

Social network analysis is a method recently used in detecting fraud. This method involves several steps:

1. It starts from modeling the relationships between major system information components (entities) as a network;
2. Suspicious components are detected on the basis of shared characteristics and there is defined a set of indicators for tracking them;
3. Suspicious entities are detected by performing simulations;
4. The resulting reports are visualized in order to be interpreted (see Fig. 5).

SAS Company, world leader in business analytics software, included SAS Social Network Analysis palette of tools for fraud detections [14].

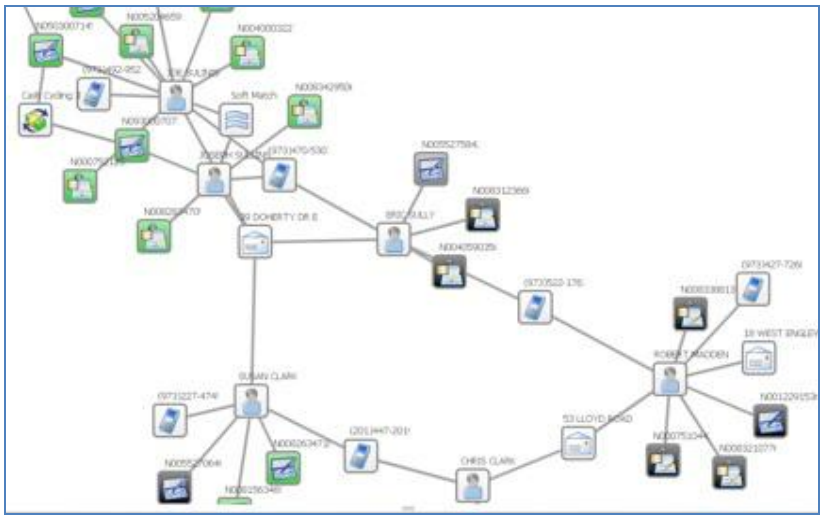


Fig. 5. Social network visualization

Social networks are linked with fraud detection because frauds are performed almost always of networks of persons rather than single individuals.

Once an individual has been identified as suspicious, the entire social network linked to him could be analyzed for searching fraud schemes. Most fraud schemes are hidden beyond the huge collections of data. If a controller would know where to look the fraud schemes would be relatively easy revealed as they are pretty simple to identify.

Conclusions

Big Data technology and distributed processing power of big data cloud bring fraud detection in insurance to another level. Not long ago, insurance fraud detection was not considered cost-effective because the cost and duration of the investigations were too high, so many

companies prefer to pay claims without investigation.

Applying Big Data analysis methods can lead to rapid detection of abnormal claims, and then creates a new set of tests to automatically narrow the segment potentially fraudulent applications or to detect new patterns of fraud, previously unknown.

The article briefly presented the National Health Insurance System and the main types of fraud that are encountered.

An analysis of Big Data technology demonstrates its huge potential, but it shows that native tools for data analysis are still immature. The analysis methods applied in the field of health insurance were briefly described, each of them being effective for a particular type of fraud or a particular stage of the fraud detection process. All this leads to the conclusion that the best solution for detecting fraud in the health insurance system is, at present, a hybrid solution, both in terms of technologies and in terms of models of analysis.

References

[1] Fawcett, T., "AI Approaches to Fraud Detection and Risk Management", Papers from the 1997 AAAI Workshop, Technical Report WS-97-07. AAAI Press;

[2] Gill, K. M., Woolley, K. A., & Gill, M., "Insurance fraud: The business as a victim", in M. Gill (Ed.), Crime at work, Vol 1. (pp. 73-82), Leicester: Perpetuity Press, 1994;

[3] Halevi, G., & Moed, H., "The evolution of big data as a research and scientific topic: overview of the literature. Research Trends", Special Issue on Big Data, 30, 3-6, 2012.

[4] Hüsemann, S., Schäfer, M., "Building Flexible eHealth Processes using Business Rules", ECEH, volume 91 of LNI, page 25-36. GI, 2006;

[5] Loshin, D., "Business Data Suited to Big Data Analytics", October 18, 2012, http://data-informed.com/business-problems-suited-to-big-data-analytics/;

[6] Melih, K., Cuneyt, A., "A Fraud Detection Approach with Data Mining in Health Insurance", Procedia - Social and Behavioral Sciences, Volume 62, 24 October 2012, Pages 989-994, ISSN 1877-0428, http://dx.doi.org/10.1016/j.sbspro.2012.09.168

[7] Palshikar, G.K., "The Hidden Truth – Frauds and Their Control: A Critical Application for Business Intelligence", Intelligent Enterprise, vol. 5, no. 9, 28 May 2002, pp. 46–51.

[8] Phua, C., Lee, V., Smith, K., & Gayler, R., "A comprehensive survey of data mining-based fraud detection research", arXiv preprint arXiv:1009.6119, 2010;

[9] Rutrell, Y., "Analytics platform helps agencies fight cyber crime, government computer news", Jul 12, 2012, http://gcn.com/articles/2012/07/12/sas-security-intelligence-platfrom-analytics.aspx;

[10] Ularu, E. G., Puican, F. C., Apostu, A., & Velicanu, M., Perspectives on Big Data and Big Data Analytics. Database Systems Journal, 3(4), 3-14, 2012;

[11] Williams, G.J., "Evolutionary Hot Spots Data Mining: An Architecture for Exploring for Interesting Discoveries", Proceedings of PAKDD99, 1999;

[12] Big Data, http://en.wikipedia.org/wiki/Big_data ;

[13] WHITE PAPER: Combating Insurance Claims Fraud- How to Recognize and Reduce Opportunistic and Organized Claims Fraud, http://support.sas.com/resources/papers/proceedings12/105573_0212.pdf ;

[14] SAS® Social Network Analysis, http://www.sas.com/offices/europe/uk/industries/banking/fraud-detection.html ;

[15] Interviu - Ministrul alternativa al

Sanatatii: 300 de mil. de euro fraudati anual - bani pentru salariile medicilor, http://www.ziare.com/politica/opozitie/ministrul-alternativa-al-sanatatii-300-de-mil-de-euro-fraudati-anual-bani-pentru-salariile-medicilor-interviu-1260060 ;

[16] 16 Top Big Data Analytics Platforms – InformationWeek http://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609 ;

[17] The IBM big data platform, http://public.dhe.ibm.com/common/ssi/ecm/en/imb14135usen/IMB14135USEN.PDF ;

[18] Big Data Debate: Will Hadoop Become Dominant Platform? http://www.informationweek.com/big-data/big-data-analytics/big-data-debate-will-hadoop-become-dominant-platform/d/d-id/1109226?

[19] Casa Nationala de Asigurari, http://www.cnas.ro/despre-noi/prezentare-generala ;

[20] Apache Hadoop http://wiki.apache.org/hadoop/PoweredBy;

[21] Ventana Research: The Challenge of Big Data Benchmark Research, 2012, http://www.ventanaresearch.com/BGD

Ana-Ramona BOLOGA (born in 1976) is associate professor at the Academy of Economic Studies from Bucharest, Economic Informatics Department. Her PhD paper was entitled "Software Agents Technology in Business Environment". Her fields of interest are: integrated information systems, information system analysis and design methodologies, and software agents.

Razvan BOLOGA (born 1976) is associate professor at the Academy of Economic Studies in Bucharest Romania. He is part of the Computer Science department and his fields of interest include information systems, knowledge management and software ecosystems. Mr. Bologa has attended over 15 conferences presenting the results of his research.

Alexandra Maria Ioana FLOREA (born 1984) has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2007 and also from the Faculty of Marketing in 2008. Since then she is a PhD candidate, studying to obtain her PhD in the field of economic informatics. At present she is assistant lecturer at the Academy of Economic Science from Bucharest, Economic Informatics Department and her fields of interest include integrated information systems, information system analysis and design methodologies and database management systems.

IBM & BDSA Collaborative Program

http://bdsa.ase.ro/ibm.html

IBM & BDSA Collaborative Program *is an initiative that aims to create a resourceful environment for students by providing them with IBM software, experienced trainers and classroom courses.*
The program is developed by IBM GDC Romania in collaboration with University Relation team and the master program ***Databases – Support for Business*** *managed by the Economic Informatics Department within Bucharest University of Economic Studies.*
These courses, held directly by Business Analytics & Optimization (BAO) practitioners, IBM GDC Romania, provide solid information in areas like: Advanced Analytics and Optimization, Business Intelligence and Performance Management, and Enterprise Information Management.

IBM's service line involved in the program, **BAO**, leverages information to enable enterprise transformation and creates sustainable differentiation through advanced information management and analytical services, deep industry and domain expertise, world class solutions required to address complex business and social opportunities across an entity's entire value chain.
The program offering consists in courses on modern business technologies, software access, course materials and experienced trainers.

Program structure
The program is focused on 4 technologies which are mapped to the master's subjects:

- InfoSphere DataStage -> Advanced Database Systems (1st year, 2nd semester)
- SPSS -> Software Tools for Data Analysis (1st year, 2nd semester)
- Cognos TM1 -> OLAP Technology (1st year, 2nd semester)
- Cognos Business Intelligence -> Executive Information Systems (2nd year, 1st semester)

Each technology is encapsulated in a separated course organized in presentations and laboratories that will held additionally to the master's subject.

Why attend the courses?

- Participation diplomas.
- Facilitated access to certification exams.
- Easier linkage to IBM's internships / job opportunities.
- Gain additional knowledge in new technologies.
- Discover IBM's professional career paths.
- Become more valuable with competitive skills.
- Learn directly from IBM's practitioners.
- Get in touch with real-life working situations.

Course schedule and enrolment
The 4-course series will start with the track InfoSphere DataStage.
Timetable:

- Applications and enrolments: 3-15 February 2014
- Course start date: 21 February 2014.

For specific information regarding the documents and actions necessary for the enrolment in the InfoSphere DataStage track: http://bdsa.ase.ro/ibm.html.

Resources
http://www-935.ibm.com/services/us/gbs/business-analytics/
http://www-935.ibm.com/services/in/gbs/bao/
http://smarterplanet.tumblr.com/
www.ibm.com/start/ro/

Contact information
Adela LUPU
Business Analytics and Optimization Consultant, Global Delivery Center Easteren Europe
E-mail: adela.lupu@ro.ibm.com