DATA SCIENCE, SPECIFIC INSTRUMENT OF KNOWLEDGE BASED ORGANIZATIONS

Claudiu Pîrnau[1], Mihail Aurel Țîțu[2], Liviu Ioan Roșca[3] and Mironela Pîrnau[4]
[1] Lumina University of South-East Europe Bucharest, claude.pyr@gmail.com
[2] Lucian Blaga University of Sibiu, mihail.titu@ulbsibiu.ro
[3] Lucian Blaga University of Sibiu, liviu.rosca@ulbsibiu.ro
[4]Titu MaiorescuUniversity of Bucharest, mironela.pirnau@utm.ro

ABSTRACT: The latest generation of database systems - the third, which appeared in the late 80s - is based on object oriented technology and is aimed for applications that store and process mainly multimedia data. In the current context, changes in the database market, leading to a significant increase in interest in those of Non-SQL type, are, in particular, due to the adoption of Big Data concept. One of the methods to be used for training the specialists in Data Science, may be authorization of an educational program in this domain having a deeply interdisciplinary character.

KEYWORDS: Big Data, Data Mining, Knowledge, Machine Learning

1. INTRODUCTION

„Big Data" term has launched a real revolution of processes, staff and technology in order to support what appears to be a new field that explodes within some giant companies like Amazon and Wal-Mart, as well as in certain bodies such as the US government and NASA, which use Big Data to achieve business tasks and/or strategic objectives. In a word, Big Data (a concept invented by Doug Laney in 2001) is the information held by a company, obtained and processed through new techniques in order to produce value in the most efficient way possible. Big Data term can be defined by the three Vs: „volume, velocity and variety." In August 2013, Mark van Rijmenam, in the article entitled „Why the 3 Vs are not enough to describe Big Data" added to the definition another 4 Vs: „veracity, variability, video and value" supported by using the following statements: „90% of all data used today have been created in the last two years. From now on, the amount of data in the world will double every two years."

2. QUESTION FORMULATION TECHNIQUE – (QFT)

Integrating Data Mining techniques (also named Called Data or Knowledge Discovery) in the knowledge based economy, has led to the achievement and development of new business models and, implicitly, to new products addressed to certain target markets. In organizations whose dimensions go beyond national borders, there are already staff having tasks within Data Mining such as Data Mining Team Manager or Data Mining Retail Analyst. Since knowledge is a combination of experiences, values, contextual information (data endowed with relevance and purpose) and intuition, the technique of wording question (generating new knowledge, ideas and questions) is classified into two categories, depending on the following: „Key components of techniques of wording questions" and respectively "Use in practice of techniques of wording questions"(for example, the development of specific procedures to knowledge transfer).Typical questions on data are „who?", „what", „where" and „when?", while questions specific to knowledge are „how?" and „why?". Questions asked before choosing a cloud solution or other analytical solution are: What are the business goals?; What architectural and management principles will be used?; What is the purpose of their implementation?; What systems current/future we rely on?; What are the performance needs?; What are the financial implications ?

A procedure needed for the transfer of knowledge can be achieved through a number of two forms F1 and F2 as follows: F1- Calculating knowledge created and shared at an organizational level, by extending the After Action Review (AAR) method, whose thinking is based on the following questions: What should happen? Why? What actually happened? Why? What is the difference? Why? What went wrong? Why? What could work better? Why? What lessons can we learn? [1]. F2 - Processing and streamlining results of an idea using SMART method, implemented based on the following set of questions: Where does this idea come? When did it occur? What were the conditions that led to its occurrence? What factors have contributed to its development? How should it be properly capitalized? Why is relevant at the moment? [2]. Based on responses, suggestions and conclusions from the use of the above questions, this type of procedure allows calculating the number of pieces of knowledge (K_{CT}) created and shared at a certain time by an organization (according to equation (1)) and the number of pieces of knowledge gained and shared from processing an idea, K_I [3].

$$K_{CT} = 6 * N_A * N_S * N_I = 6 * 4 * 24 * 30 = 17280 \text{ [knowledge]} \qquad (1)$$

Where: number of questions used in the method AAR=6; number of subtasks analysed = NA;

Number of suggestions/conclusions/results = NS; number of participants =NI.

An important element in this calculation methodology is the applicability rating of an idea results, RA, shown in Table 1.

Table 1. Applicability rating of an idea results

Level of geographical dispersion	Rating	Applicability
Local	1	Restricted
Regional	2	Extended
National	3	Large
International	4	Very large

In this situation, the number of pieces of knowledge accumulated and shared as a result of processing an idea (K_I) can be calculated with the help of equation (2).

$$K_I = R_A/6 * (N_S + N_D + N_I + N_Q) \qquad (2)$$

Where: R_A = Applicability rating; 6 = Number of questions used by SMART method; N_S = Number of suggestions generated by the answers to SMART method; N_D = Number of sectors benefiting from the implementation of the idea; N_I = Number of new ideas generated; N_Q = Number of new questions generated.

The main questions Knowledge Manager need to ask himself in the process of knowledge management (KM) integration in the smart sustainable development of organizations are: Where do you start in applying KM?; Is KM better implemented bottoms-up or top-down?; What are the difficulties or challenges of KM?; What will ensure the success of KM in an organization?; What is the “knowledge cycle”?; Is there any personal benefit from KM? [4].

The main questions related to research design and data analysis, known as „Crotty's model” are: What knowledge claims are being made by the researcher (including a theoretical perspective)? What strategies of inquiry will inform the procedures? What methods of data collection and analysis will be used? [5]. In the process of preparing future professionals in Data Science, we will also use many answers and ideas obtained after several series of important questions.

3. SPECIFIC EDUCATION IN DATA SCIENCE

An education program specific to Data Science could be established based on the following structure:Research design and data analysis, Exploration and data analysis, Applied Machine Learning - (AML), Visualization and data communication, Developing team-working competences. **Research design and data analysis**, based on the use of different research methods, technique of wording questions, intelligence management etc., as follows:

- Research design, cantered on using *quantitative research methods* based on collection and processing of data through experiments, surveys, observations etc., highlighting certain results of data processing, namely the use of *qualitative research methods* (as a branch of empirical research), which refer to the attributes of a person or a group of people who can be described through motivations, aspirations, attitudes, values, culture, lifestyle, behavior that are rendered as accurately as possible. [6]
- Technique of wording question - can be divided into two chapters, as follows: Key Components of the Question Formulation Technique & Experiencing the Question Formulation Technique;[7]
- Data analysis (Management Intelligence) and decision-making processes. Data quality analysis involves completing of five stages: setting goals of analysis of data quality and structure of collected data; preliminary data review; selecting statistical test; checking forecast in statistical test; determination of conclusions. [8] The study of decision-making processes involve analysis of planning, organizing, training and control functions. An important role should be also granted to the study of decision making typology (autocratic, consultative group) and optimization techniques of decision-making process;
- Analysis of cognitive biases, which involves (cold, hot) analysis of various forms of prejudices and their links with domains such as psychology and behavioral economics;
- Direct and indirect digital influence on actions, opinions and human behavior centered on the role of influence-providers (Example: social commerce) in the marketing process;
- Integration of Data Mining in the knowledge based economy;
- In knowledge-based organizations, the training in this domain is ensured through a “Senior Data Visualization Specialist".

3.1 Exploration and data analysis

Process based on research design through mixed methods (quantitative and respectively qualitative) and statistical analysis (measurements, inferential statistics and causal inference). Statistical techniques and methods of using the R language (implementation of S language in the open-source environment). Topics covered in quantitative techniques include: descriptive and inferential statistics, sampling, experimental design, parametric and non-parametric tests of difference, regression squares and logistic regression.

Storing and data retrieving, based on analytical applications, Big Data, distributed data processing, the use of relational databases etc., as follows: **Analytical applications (Big Analytics) and cloud solutions for Big Data**. Design of solutions should take into account mainly of four elements: data sources have different scales (multi-terabyte or arena petabyte); Speed is a critical element (use of Extract Transform Load- use LTE technology may not be sufficient, requiring solutions like S4 or STORM); Storing patterns are changing (in order to store unstructured data we can use the Hadoop Distributed File System or Amazon S3); It is necessary to support multiple analysis paradigms and calculation methods used, such as analysis of databases in real-time, interactive queries of data warehousing distributed by massively using parallel processing and distributed processing engines based on simple (aggregation) algorithms or complex ones (machine learning); **Computational solutions in the process of storing and retrieving data.** The importance of information infrastructure (US). Hybrid Storage Systems (IBM Smart Cloud Virtual Storage Centre); **Distributed data processing; Using relational database; Using Graph Databases model** - the graph is a set of objects (nodes) and relations between them (edges) - in the context of growing hybrid databases role (SQL + Non SQL) within Big Data; **Streaming Data for Big Data,** initiated on the question "What are the data that are not at rest ?" The answer would be, the systems that manage active operations, and therefore should have persistence. In these cases, the data will be stored in an operational data warehouse. However, in other situations, these operations were executed, and it is time to commonly analyze the data in a data warehouse or Data Mart. This means that information is processed in batches and not in real time. Streaming Data is an analytic calculation platform focused on speed. This is due to the fact that these applications require a continuous flow of unstructured data, which need to be processed ("time windows"). Therefore, the data are continuously analyzed and transformed into memory (using cluster type servers) before being stored on a disc; **Cube Technology** - development and implementation of some innovative solutions for websites that manage events (Example: organizing of modular stands) and buildings, including time management and optimization of working conditions. A practical application in this regard is the KELIPSE solution based on terminals for modular stands [9]. Another category of applications can be considered achieving cloud private solutions, public or hybrid, by Cube Technology Business IT Solutions London Company [10].

3.2 Applied Machine Learning - AML

AML is a discipline that focuses on developing algorithms. The relationship between data and predictions/models is taught (using automatic machine learning techniques) by examining a substantial amount of relevant information. The focus is centered on intuition and practical examples and less on theoretical results. It is good that participants have minimum skills in probability, statistics and linear algebra. The main chapters of this discipline, could be considered the following [11]: **Experimental Design** - planning an experiment is based on three elements: an idea, justification (possibly by developing a hypothesis) and prior documentation in specialized literature. Designing the experiment should start from the assumption that it is based on commands and methods that can be verified and measured. The existing variables/outcomes in an experiment can be classified into four categories: variables (the amount does not have a fixed value); independent variables (defined by researcher), dependent variables (measured in the experiment) and external variables (resulting from uncontrollable variations/deviations) that may affect the results of the experiment. Examples: sensory tests, sensory characteristics, standardized criteria by means of an experimental protocol; **Learning algorithms** - the list of the most common learning algorithms, includes: linear regression, logistic regression, decision trees, algorithms/techniques for reducing the dimensionality; **Engineering elements specific to AML**. Designing applications Machine Learning type requires two components: understanding the properties of the task to be solved and limitations of the model that will use the application. Typical cycle of a Machine Learning type application consists of four stages: design a set of characteristics (pattern/template type or combined), running the experiment and analysis of results (based on a data set), modify the set of functions used and return to the design stage. Limits of application, including running speed, are influenced by the amount of input data; **The predictions and interpretations issue**. The problem of merger between explanation/interpretation and empirical prediction is already a common element, but the distinction (recognized in the philosophy of science) should be understood in view of scientific progress. The purpose of this chapter is to clarify the distinction between explanatory and predictive modelling, to discuss about its sources, and to reveal the practical implications of the step by step distinction process modelling; **Network analysis** - In general, network analysis is a structured technique used to analyze mathematically a circuit or a network of interconnected components. **Collaborative filtering. Recommendation systems with collaborative filtering**. *Example*: We can determine a user's profile based on his uploaded photos, favorite photos, friends, groups to which he participates (a data set). For a picture, we can infer its contents using the list of associated tags. A considerable step forward would be the use of image analysis and object recognition procedures. The next step is to design recommendation mechanisms that can be applied based on these data. There are two major classes used in recommendation, namely: collaborative filtering and content-based recommendations. For a visitor (unregistered user), not having a profile, we can only examine the content of the visualized image and guess what he would like to see further in this way: **recommendations based on content**. For **a registered user**, because we know his profile, we can apply collaborative filtering to discover users with similar profiles. Starting from these users, we can get pictures that we may recommend (through techniques such as selection from their favorite photos sets): **recommendations based on profile**. For **a registered user viewing an image,** we can achieve an hybrid algorithm: on the one hand we seek users similar to him, on the other hand, we use information related to the contents of the image viewed to refine recommendations: **recommendations based on "context" (user profile & image content viewed)**. Subsequently, we can establish a set of functional requirements, modelled by **use cases diagrams**, then we set the sequence of actions, modelled by **sequences diagrams** [12].

3.3 Visualization and data communication

Visualization and data communication focused on design and implementation of complementary visual and verbal representations, based on analysis of patterns in order to convey findings and answer to necessary questions of decision-making process, and to provide conclusive evidence, supported by data. The minimum necessary elements that can be used in this educational process includes, among others, exploratory data analysis, the effective visual presentation of data, etc., as follows:**Exploratory data analysis** studies the existence of different types of data, descriptive statistics, graphical representation of a set of data, examination of variable distributions, production and use of statistical tests, etc. From the regressive models used in this case, we can mention: correlation coefficient, matrix of correlations, spreading diagram. Linear regression. Non-linear regression (polynomial and mixed exponential). Multi-linear regression. Logistic regression. Notions of survival analysis. Cox's proportional hazards model. Additive models. Temporal series integrate elements such as: smoothing methods. Prognosis using the trend. Prognosis using the trend and seasonal component. Dynamic models based on temporary series: explanatory models, adjusting models, auto-predictive models, ARIMA model. Among clustering algorithms, used for data mining and identify natural groups, we can mention k-means partition algorithm (supporting text mining and clustering techniques based on hierarchical methods)and detection algorithm of anomaly (analyses the characteristics of normal cases to signal unusual cases);**Effective communication in writing** follows a number of three stages: Stage of preparation of written communication (Communication purpose, Audience identification, Setting the main idea, Setting the format). Drafting stage of written communication (Drawing up the text, Reviewing, Editing). Questions regarding the improvement of written expression skills (related to structure, style and content of written communication). **Effective visual presentation of data** – the type of stored data has major implications on how they will be presented. **Nominal data** are discrete and have no intrinsic order (gender, race, etc.). **Common data** have a prescribed order, such as the level of satisfaction (very satisfied, dissatisfied), size that fits (small, medium, large) or ownership status (owner, tenant). Display in a different order would not make sense as they are not really numeric. **The data range** consists of a series of sequential numerical ranges, which has a distinct order and can be divided into equal parts or can be sorted in ascending or descending order. In this category can be assigned data referring on time (months of the year), age, salary and other financial measures. **The main phases of an efficient data view, are**: Be open to discovering new insights; Think big but start small; Design for your user; Prototype to identify needs (see figure 1.); Obtain feedback early and often [13]. **Adapting design on visualization and communication of data, according to human perception (Design for human perception) -** In this discipline can be studied, from the point of view of computer science, the limits of human perception based on the following steps [14]:**Identifying relationships between perceptual psychology and the science of image**. The studied items will help in getting effective answers to questions such as: How should colors be used? What graphical entities can be measured accurately? How many distinct entities can be used without creating confusion? These questions will be integrated in *Visualization design*. Under *Computer Vision*, will be sought answers to questions such as: What primitives people cannot detected in due time? What level of accuracy is perceived through various primitives? How can we combine primitives in order to recognize complex phenomena?**Information theory and human perception**, in which, for each (visual, acoustic, etc.) primitive can be effectively measured the number of distinct levels that an individual can identify with a high degree of accuracy. Each level will be labelled according to "channel capacity" of transferring human information that will subsequently be measured in bits. Then, will be studied **the main reasons related to one-dimensional stimuli** (sound, salinity, vibrations, etc.) and **absolute judgements concerning the multidimensional stimuli**, such as salinity and sweetness, dimension and brightness or multiple parameters related to sound: frequency, intensity, discontinuation rate, duration and location. These judgements are in accordance with linguistic theory, which identifies 8 to 10 dimensions, where every distinction can be of binary or ternary type. **Comparisons between the concepts of measurement and detection** will allow measuring the distance through a mathematical calculus of difference type (absolute value), different from the method of detection (relative value). In practice can be made the difference between a total of 10 tasks related to graphic perception: angle zone (area), color, saturation, density (percentage of black color used), length, the position on a scale, inclination (slope) and volume. **The study of persuasive communication in advertising**, based on *concentration on the message and final satisfaction expectation by the receiver*, depending on his capabilities, enables analysis of the decisions taken following absolute judgements (based on the amount of information), the range of perceptual dimensionality and modalities of reconfiguration of absolute judgements sequences (resulting in analysis of immediate memory, based on the amount of information, regardless of their complexity). Human capacity on the distinction between absolute levels of information is limited to a number of 4-7 values.

Developing team-working competences, necessary to every human individual can be achieved by means of four chapters: training and groups preparation, groups management, training sessions for groups and, respectively, case studies. The study of this discipline implies, among other things, **analyze of key dimensions of a team, assessment of main components of team effectiveness, development of cross management** (oriented towards establishing relationships and less towards defence of domains in functional area, involving the realization of distinct projects by working teams within a department or a company), implementing and developing the concept of empowerment (this involves primarily granting freedom to each employee to contribute in the process of decision-taking, distribute power at the level of each employee according to his competence and in accordance with the objectives and culture of the organization), as well as **analysis/development of models of virtual teams** [15-18].

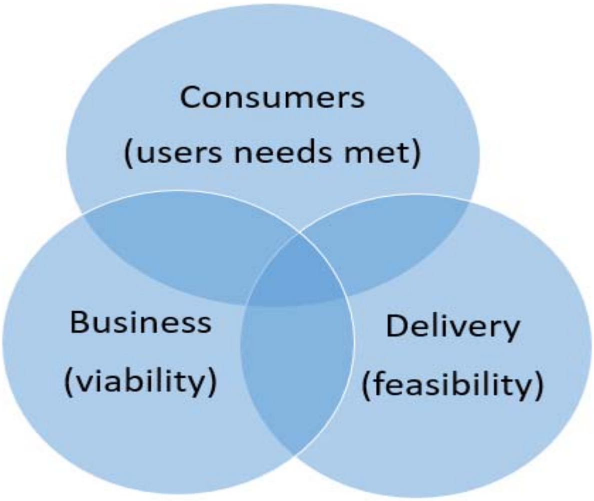

Figure 1. Prototyping and evaluating ideas

4. SOCIAL COMMERCE IN KNOWLEDGE BASED ECONOMY

Development and promotion of social commerce as influence factor of marketing processes and objective of knowledge based organizations ("Markets means conversation" - a concept promoted by "Amazon" and "E-bay"), is a derivative of electronic commerce that involves using forms of "social Media "and on-line content in order to foster social interaction and users contribution. This approach should support buying and selling of products/services in social environment. [19] The merit of this new game of marketing consists in the art of conversation and the power to persuade the customer through the message sent. The main player in the social commerce market is "Facebook.com" socialization website with over 600 million global users, representing 10% of world population, its upward trend is still maintained. Social commerce consists of the following elements [20]: Buyers community (GDGT); Community of group buyers (Groupon); Share information on purchases (Just Bought It); Purchasing products (Polyvore, Pinteres); Social advice (Fashism); Co-shopping (specialized search engines, such as "Shop Together"). Social commerce pillars, six in number, allow setting of new trends in the knowledge-based economy (see Figure 2), including within strategic alliances of cluster type: Visibility - social networks are the ideal environment for the presentation of new sustainable products; Reputation - the existence of a brand image and sustainable identity; Proximity - allows shortening the distance between brand/product and potential customers; Contextualization - sustainable products/services must reach the right place, at the right time and the right people; Recommendation - the existence of social platforms that can support/promote strategies of the organization. Purchase of sustainable products/services is based on recommendations/references from the part of acquaintances; Customer support - existence of a suitable space within which we can demonstrate customer care.

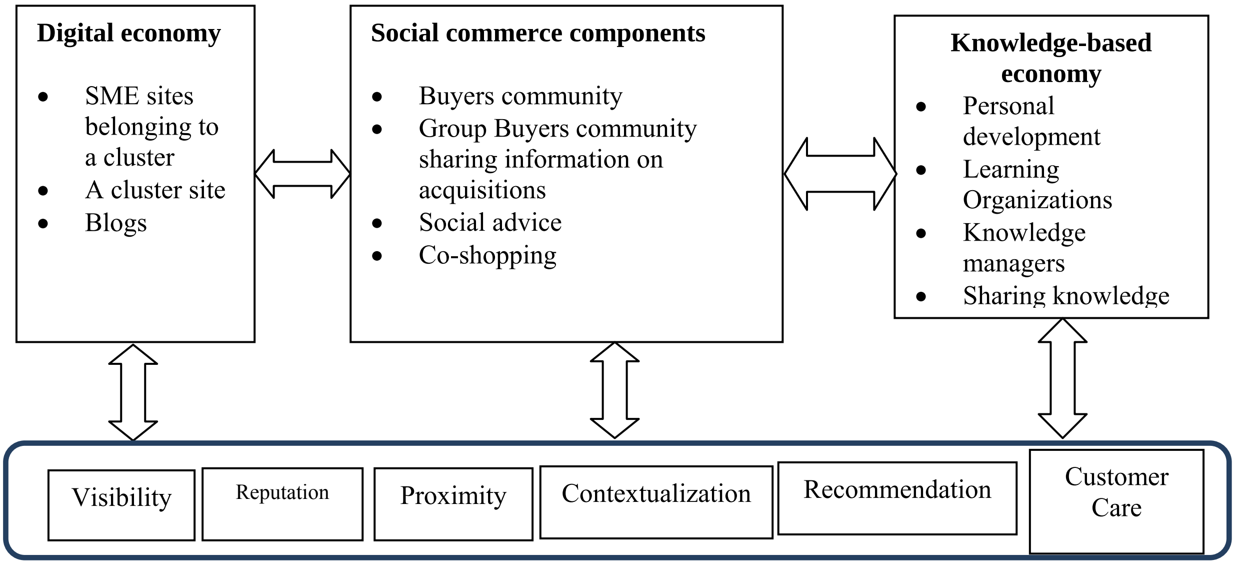

Figure 2. The role of social commerce in knowledge-based organizations

According to statistics, the turnover generated by social commerce has reached over $30 billion in 2015 [21]. Social commerce will become a key lever of the purchase act, based on sales through social networks and on influencing (adaptation) consumers through advice and experiences shared among them.

5. CONCLUSION

Development of social commerce is a derivative of electronic commerce that involves using forms of "Social Media" and on-line content in order to foster social network and users contribution. Social commerce will become a key lever of the purchase act, based on sales through social networks and on influencing (adaptation) consumers through advice and experiences shared among them. Fundamentals of effective data visualization and communication involve understanding visual elements aligned to pre-cognitive thinking, to the modality of selection of suitable communication channel depending on the target group, and positioning visualization of data in context, as an effective communication tool.

6. REFERENCES

1. *** *The After Action Review,* Mission-Cantered Solutions Inc., Colorado, USA, (2008).
2. Kempf, K.G. Keskinocak, P. Uzsoy, R. *Planning Production and Inventories in the Extended Enterprise,* Springer International Publisher, (2011).
3. Pîrnău, C. Contributions On Integration Of Knowledge Management In The Sustainable Development Of Small And Medium - Sized Enterprises, PhD Thesis, Lucian Blaga University of Sibiu, March (2015).
4. *** Frequently Asked Questions on Knowledge Management - CCLFI. Philippines 2009;
5. Creswell, J.W. *Research Design. Qualitative, Quantitative and Mixed Methods Approaches. Second Edition.* Sage Publications, International Educational and Professional Publisher, California, USA, (2003).
6. Bîrsan, M. *Research Methodology. Lecture Notes.* Centre for European Studies, Alexandru Ioan Cuza University of Iasi, (2012).
7. Day, J. *Experiencing the Question Formulation Technique (QFT^{TM}),* The Right Question Institute. A Catalyst for Micro-democracy, Cambridge, Massachusetts, (2014).
8. Ciora, L. Buligiu, I. *Methods and techniques to analysis data quality*, Economic Informatics Magazine, pp. 59-64, No. 1(25)/2003, Romania, (2003).
9. www.cube-technologies.com/fr/.
10. www.cubetechnology.co.uk.
11. Brodley, C.E. Rebbapragada, U. Small, K. Wallace, B.C. *Challenges and Opportunities in Applied Machine Learning*, Association for the Advancement of Artificial Intelligence, pp. 11-24, ISSN 0738-4602, USA, (2012).
12. Trausan-Matu, S. Popescu, A.E. *Collaborative filtering recommender systems*, The Faculty of Automatic Control and Computer Science, Polytechnic University of Bucharest, (2016).
13. Luu, L. Design thinking & big data analytics. Five principles for Effective Data Visualizations, Thought Works Inc., USA, (2016).
14. Rosenholtz, R. Dorai, A. Freeman, R. *Do Predictions of Visual Perception Aid Design?,* National Science Foundation, Division of Behavioural and Cognitive Sciences, Grant 0518157, USA, (2005-2009)
15. Lonchamp, J. *Collaboration Flow Management: a New Paradigm for Virtual Team Support*, Inria Ecoo Project, France, (2003-2007).
16. Iacob (Ciobanu), N.M.*Distributed Transactions in Transnational Companies*, Annals of Ovidius University of Constanța, Economic Sciences Series, Vol. XI, Issue 1, pp. 963-966, (2011).
17. Iacob (Ciobanu), N.M. *The Distributed Transaction Management in the Modern Economy*, Annals. Economics Science Series, Tibiscus University of Timişoara, Vol. XVII, pp. 739-744, (2011).
18. Țîțu, M.A. Pîrnău, C. Pîrnău, M. *Creativity, Education and Quality for Sustainable Development, the real Support for the Innovative Cluster's European Network*, 8th Research/Expert Conference with International Participations, „QUALITY 2013", Neum, Bosnia & Herzegovina, June 2013, pp. 19-24, ISSN 1512-9268, (2013).
19. Gay, R. *Online marketing*, Oxford University Press, (2007).
20. Grossek, G. *Internet Marketing Communications*, Lumen PH, Iasi, (2006).
21. www.culturecrossmedia.com/social-commerce-est-ce-lavenir-du-e-commerce-2/.