THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

DATABASE SYSTEMS JOURNAL

Vol. VI, Issue 2/2015

LISTED IN

RePEc, EBSCO, DOAJ, Open J-Gate,
Cabell's Directories of Publishing Opportunities,
Index Copernicus, Google Scholar,
Directory of Science, Cite Factor,
Electronic Journals Library

BUSINESS INTELLIGENCE

ERP

DATA MINING

DATA WAREHOUSE

DATABASE

ISSN: 2069 – 3230
dbjournal.ro

Contact

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

CONTENTS

Architectures for the Development of the National Interoperability Framework in Romania

Codrin-Florentin NISIOIU
Bucharest University of Economic Studies, Bucharest, Romania
codrin.nisioiu@ie.ase.ro

The authors of Digital Agenda consider that Europe do not take fully advantage of interoperability. They believe that we need effective interoperability between IT products and services to build a truly Digital Society. The Digital Agenda can only be effective if all the elements and applications are interoperable and based on open standards and platforms. In this context, I propose in this article a specific architecture for developing Romanian National Interoperability framework.
Keywords: *interoperability, collaborative working environment, cloud computing*

1 The Interoperability dimensions from the European Interoperability Framework.

Political context. The political support for achieving interoperability is an absolute necessity. For action cooperation to be effective in the achievement of the common objectives, it is necessary that partners share common visions, focus their efforts and resources in the same directions, use the same timeframe and synchronize their changes determined by mutual agreement.

In the European context, the political support for the achievement of interoperability can be reflected by the specific political instruments, such as the European Directives, ministerial statements and multiannual programs. These instruments express the vision and the priorities of the European policy makers, in whole or in part. The level of funding, the budgetary issues, the measures and deadlines imposed can offer additional details about the political priorities and the understanding of the political context.

An important challenge in the context of political changes in the European Union is the management of the cross-border services development in order to ensure their continuous development and support. Namely, the challenges are: avoidance and/or prevention of divergences in the vision of interoperability and the insufficient support in the member states. The best way to ensure a continuous support is through the ongoing activities of the various bodies of coordination and consultation, especially any permanent structures dealing specifically with the interoperability problems.

Legislative aspects. The interoperability requires proper timing of the legislation in the Member States which are cooperating so that the electronic data from any of the Member States can be in accordance with the law and recognized wherever it is necessary to be used in any other Member State.

Speaking of the legislative aspects of the interoperability, these are necessary for a variety of reasons, especially for: mutual recognition of electronic data from other Member States of the Union, mutual assistance for process integration and cross-border processes through competent institutions of the Member States. A solution for the legislative problems which are related to the lack of legislative clarity of electronic data protection is the implementation of some pilot systems for several Member States. Afterwards, their example may be followed by other countries. Thru these pilot systems the market entry barriers are reduced, eliminating the need to solve conflicts and other problems that may result from the compliance ([9,10]) of the 27 sets of constraints.

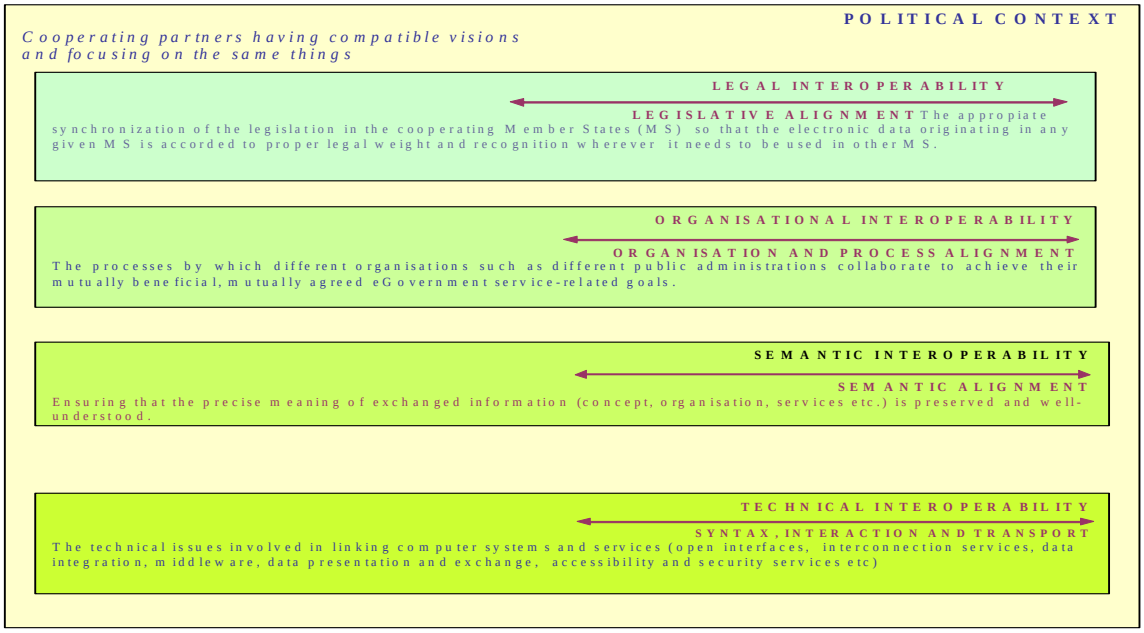

Fig. 1. The interoperability dimentions from the European Framework of Interoperability, version 2.0.

The data protection in the cross-border context is one of the key legal issues. The question that arises here is whether there is enough legal and operational support to cover the entities and the mechanisms which are responsible for ensuring data protection. The answer to this question may be provided by a data protection strategy, which should include elements such as the designation of one or more data protection authorities and a planning for the establishment of some collaborative structures and of the associated mechanisms. The Commission and the Member States should assess the impact of the legislative proposals on the ICT and the interoperability should be included as a standard criterion in the public procurement process being preferred the choice of some open standards and specifications.

Organizational Interoperability - The organizational interoperability allows the definition of the business objectives, the business process modeling and the collaboration of the administrations which want to exchange information and have internal structures and different business processes. The organizational interoperability addresses to the users requirements through the implementation of the basic electronic services, making them easily identifiable and user-centered. For a better approach between public administration and citizens or companies, the Member States use the important events in the lives of citizens (birth, marriage, death, etc.) and the business stages for the companies (the setting up of a business, the liquidation, etc.) by providing them under the form of basic services by electronic means. In this way citizens and companies remain focused on their needs and do not have to focus their efforts on the understanding of the functional organization specific to the public sector. The provision of the services is transparent and customer oriented.

Each of the life time events and of the stages of business is associated with the relevant actions and with interactions with and among public institutions. Electronic services may involve one or more business processes to be performed in a given sequence between different administrations.

The cross-border services should be determined jointly by the participating administrations through a demand-driven approach, but the responsibility should be decentralized.Decentralized responsibility involves the ability of each partner to organize its business processes in a manner best suited to its national practices. It is

unrealistic to believe that the administrations of different Member States will be able to harmonize business processes due to the cross-border requirements. The stages and the internal processes of a Member State may remain unchanged only if the "points of entry" and "exit" to these processes are made transparent and interoperable to other Member States which are involved.

The public institutions that are providing cross-border electronic services should analyze the business processes involved and to agree on interoperable interfaces Business (IIB) and on their specific standards (Fig. 2.).

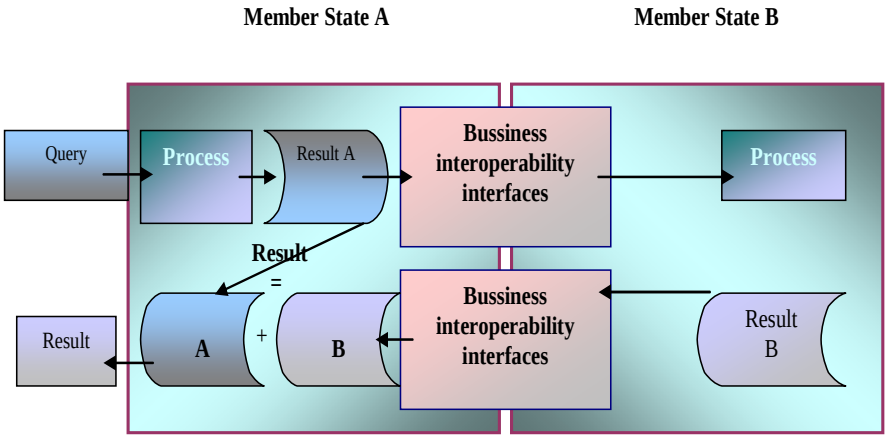

Fig. 2. Interoperable interfaces of business

IIB application is permitted if the administrations concerned have agreed in advance which are the cross-border services that are interested in, which business processes are involved in this, which administration provides the necessary functionalities for the interconnection between administrations through IBB of the national business processes (which may be completely different at the organizational, semantic and technical level).

To operationalize the cross-border processes there are necessary the following steps: the alignment of the business processes, the business process reengineering, the establishment of the service level agreements, the assessment and the dealing of the gaps, the change management, the enhanced collaboration.

The entities (Member States, administrations) involved in the alignment of the business processes must align their used standards for describing business processes. A collection of business processes and the best practices in the Member States concerned is absolutely necessary because it allows the reuse of the best practices between Member States.

The business process reengineering is an interim solution to achieve organizational interoperability necessary to provide cross-border services. To achieve this it is necessary a cross-border effort to analyze the used business process, aimed at a common understanding of the business processes, identifying common elements and the decomposition of the process in the processes components in a way to allow the cross-border interconnection.

The establishments of agreements at service level allow formalizing specific aspects of the mutual assistance, joint activities, business processes "coupled" in order to provide cross-border services. One of the means is the memoranda of understanding between governments, detailing bilateral agreements on joint actions and cooperation. We consider the establishment of service level agreements as a standardizing cross-border activity, where the standards which have to be defined and implemented are exactly these instruments.

The assessments of the common assessment framework should be carried out at the sectorial level, in order to identify the real deficiencies of the business processes. The identification of the deficiencies allows the improving and the alignment of the business processes.

In order to set up the change management, the Member States must establish a change management strategy at national level and to integrate it in the action plans for achieving cross-border services.

The strengthening cooperation of the Member States is provided through:

- ✓ the cross-border exchange of information on business processes;
- ✓ the cross-border consultations about the taxonomy of business processes and its components;
- ✓ the cross-border coordination of the change management activities;
- ✓ the cross-border functional and sectorial coordination;
- ✓ the cross-border assessment of the sectorial deficiencies that affect specific activities electronic services;
- ✓ The cross-border consultation on the mechanisms and the orchestration architecture for the cross-border business process.

Semantic interoperability - Semantic interoperability allows to all application to understand the exchanged data and allows also to the systems to combine information and the resources in order to process them in a meaningful manner. In practice, this will involve the establishment of common sets of data structures, data and protocols. The data which is to be exchanged may become interoperable if the responsible administrations:

- ✓ publish the information on the data involved at national level;
- ✓ agree on data and data dictionaries needed across borders;
- ✓ agree on multilateral mailing lists between different national and cross-border data.

The essential requirement for information exchange is the existence of a single language that allows the description of the meaning and the basic structure of data involved. The development of a common XML semantics must be done in a coordinated manner and it should be considered in cooperation with existing standardization bodies. The definitions and the European schemes should be made available to interested parties (stakeholders) through a common infrastructure. The Semic.eu portal aims to lay the foundations of semantic interoperability necessary for cross-border services in all the activity sectors and at all levels both conceptual as well as implementation.

The European Commission and the Member States should identify and support the development of the sectorial communities whose role is to enable semantic interoperability. The sectorial communities are the entities that have the best knowledge about the reference models, the services they use or offer and also the problems they are facing. The knowledge and the expertise of the sectorial communities' members should be focused on standardization efforts.

The national interoperability frameworks should take into account the transboundary nature of the semantic interoperability when the data dictionaries are developed.

Technical interoperability - Technical interoperability includes the key issues for connecting the systems and the services through open interfaces, through the interconnectivity of the services, through the interconnectivity of the data, through the exchange and the presentation of data, through the accessibility and secure services. Technical interoperability should be applied both at the front-office level (Fig. 3.) and at the back office level of the system (Fig. 4.).

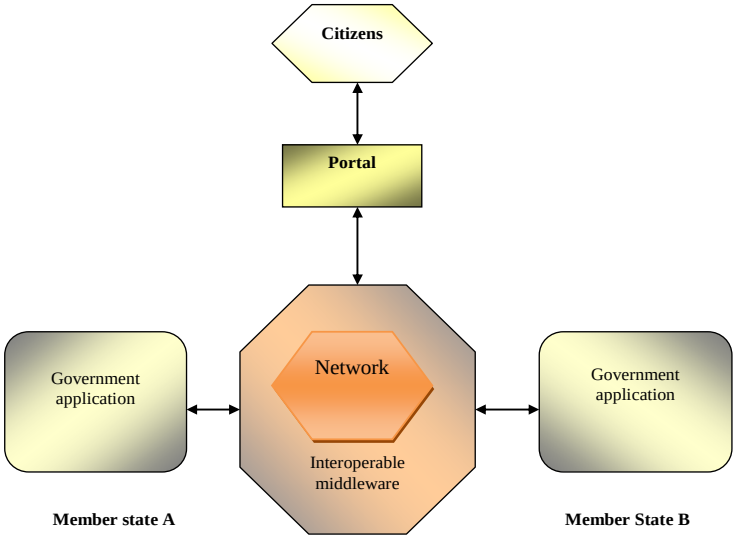

Fig. 3. The complex interactions through a portal

The aspects that have to be considered at front office in order to achieve the technical interoperability are:

- ✓ the exchange and the presentation of data;
- ✓ the availability – the design principles of interfaces;
- ✓ the multichannel access;
- ✓ the character sets;
- ✓ the file types and the documents format;
- ✓ the file compression.

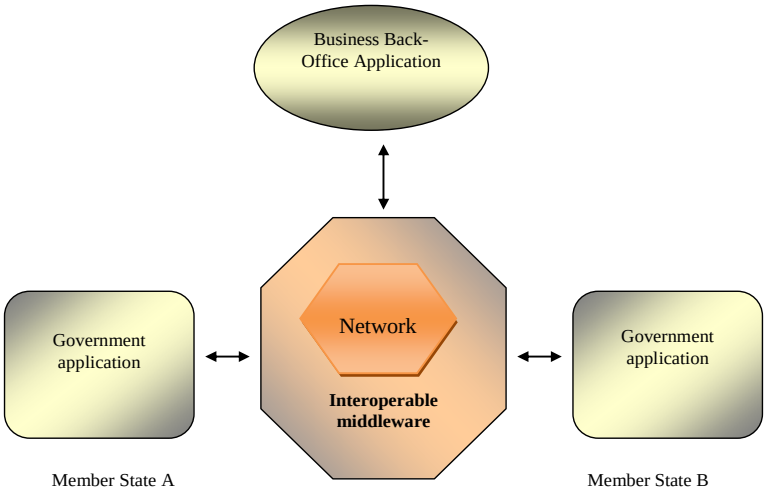

Fig. 4. Interactions in the middleware

The aspects that have to be considered at back-office in order to achieve the technical interoperability with the business applications from the back-office are:

- ✓ the integration of data;
- ✓ the XML-based standards;
- ✓ the EDI-based standards;
- ✓ the web services;
- ✓ the architecture of distributed applications;
- ✓ the interconnection of services;
- ✓ the transfer protocols of the messages and of the files;
- ✓ the security and the transport of messages;

- ✓ the messages storage services;
- ✓ the access to email;
- ✓ the services directory and type name;
- ✓ network services.

The European Union and the Member States administrations must have a clear and a precise image on the used technologies, on the technical expertise and the capacity of their staff and on the documentation of business processes. The administrations must also engage themselves in auditing, compliance and benchmarking to identify closed systems and other barriers in order to obtain technical interoperability.

Analyzing the information provided by the National Interoperability Frameworks observer (NIFO) [11], there is a list of the mature interoperability frameworks comprising the following countries: Bulgaria, Denmark, Estonia, Germany, Greece, Hungary, Italy, Poland, UK and Switzerland. Following a web analysis, there were eliminated from that list the frameworks that do not support an international language (English, French) and are not at least the second version. The new list includes Bulgaria, Estonia, Germany and the UK.

The specific frameworks from Bulgaria and Estonia provides general directions of their development and implementation without a detailed presentation, while Germany and the UK provides a detailed overview of the general directions of development and implementation, proposing solutions that can be integrated in other national interoperability frameworks.

Following these examples, the recommendation for the creation of the National Interoperability Framework in Romania is that this should be made in collaboration with Germany and UK which would increase its development through the transfer of know-how, best practice examples that can be provided by these countries.

2. National Interoperability Framework in Romania

A possible architecture for a collaborative solution [4,5,6] in order to develop, update and maintain the national interoperability framework in Romania is shown in Fig. 5.

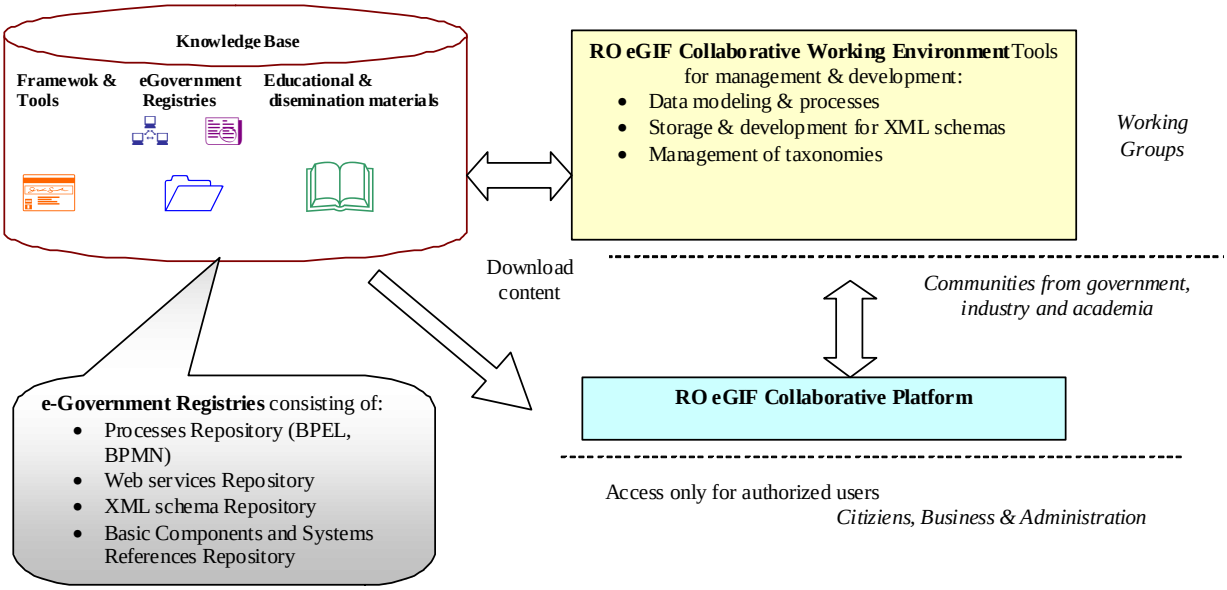

Fig. 5. The collaborative solution architecture for developing, updating and maintaining the National Interoperability Framework of Romania (ROeGIF)

The collaborative platform should contain at least 4 components: a content management system made by the working groups with the purpose to disseminate research conducted by multichannel means, a system of training management related to the specific e-skills, an open social platform and a multilingual library of case studies in the field of e-Government, with the purpose to assess e-competences acquired through them.
The components of the collaborative platform must be interconnected.
To implement the proposed collaborative architecture it is necessary to use as many solutions (open source) and open standards as possible. They allow a better channeling towards the development of the required e-government components, the reuse and the improvement of the existing ones and, not least, the existence of a community of developers that can allow the further development of the prototypes proposed in our architecture. It should be also implemented a collaborative working environment using a private cloud architecture.

3. Collaborative working environment in the cloud

The architecture will be analyzed from two perspectives: one is for the cloud functional services and the other is for the systems involved in the cloud.
From the perspective of the functional services, the services and functionalities offered by cloud respond to the user's need which initiates a request for services or other resources through the self-service portal. The cloud applications are seeking for resources that meet the requirements using the cloud services portfolio. The access to resources and services is provided on through the portal too. The services are provided to the user by the administrator (cloud service provider). The administrator uses the management portal in order to make public the services that are offered through the cloud and to prepare resources to be used in the cloud. The portal represents an interface to a range of tools and reports, as it can be seen in Fig. 6.
The cloud services provided are: the services for business processes and information, services for software platforms and infrastructure services.
The services for business processes focus on providing existing business processes through cloud community. If the steps of an existing process are known, this can be provided as a service via the service catalog. This allows to the cloud services provider to automate any stage of the process while leaving the changes transparent to the client. Information can be provided as a service, thus enabling the organization to improve the relevance and to streamline costs for information. Information becomes available to the people, to the processes and to the applications, thereby improving operational impact.
The services for the software platforms allow the user to select an instance of particular software that is intended to be created without the need to know where and how it will be hosted. Such type of services includes workflows that enable the design, the development, the implementation and the hosting, and also the collaboration, the integration, the database integration, the security, the scalability, the storage, the persistence, the state management, the versioning applications, the orchestration and the possibility of developing a community of developers. These services are provided in the form of integrated web solutions. The key components of the services for software platforms include tools and services for developers, for using the software in a dynamic manner, for reporting and for optimized middleware - application servers, database servers and portal servers.
The infrastructure services enable the provision of standardized computing resources. They allow the user to request and receive the instance of a computer system without the need to focus on IT issues such as the placement in the network and hardware availability.

The Cloud Management Platform includes: business support systems, service delivery platforms, tools for defining, publishing, analyzing and reporting services.

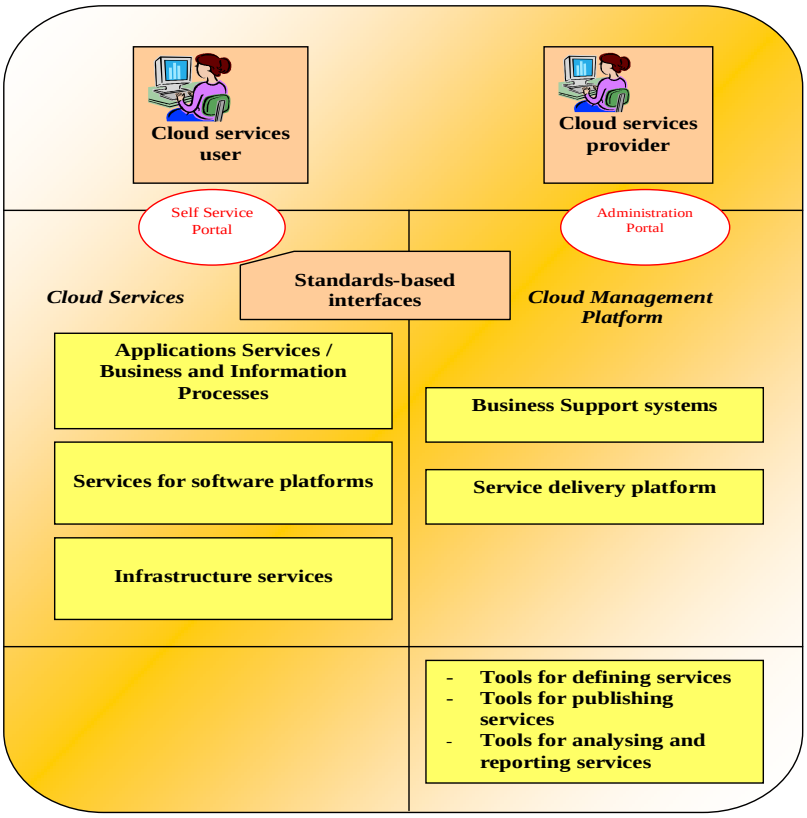

Fig. 6. The proposed private cloud architecture - the perspective of functional services

The business support systems are designed to support the cloud services provider in taking orders, processing invoices and collecting payments. The services delivery platform enables rapid development and the implementation of services.

The tools for defining services allow the administrator to create and to modify the services offer that will be made available to the user, to define what information is needed in order to meet the user's demand, what are the necessary automated activities and manuals to meet this demand and what resources from the datacenter will be used. In addition, the administrator can establish the services level agreements and the costs associated with the entire request.

The tools for publishing the services allow the administrator to publish and to eliminate the types of requests from the services catalog. The administrator can determine who can access the service and who may require a service. The tools for analyzing and reporting the services provide performance reporting to the owners of services and to their applicants.

A private cloud is composed of management environments and managed environments, as can be seen in Fig. 7.

The environment for the management of cloud services supports services throughout their life cycle. This layer acts as the control center that effectively manages the entire cloud environment. The combination between the environment for the services management and managed environment by cloud ensures that resources are efficiently managed from the data center and can be predicted, provided and configured quickly. The products and the services that perform automated management help in defining and managing the supply of services associated

with hardware resources throughout the life cycle of the service. Automated management of services usually includes: the catalog of services, the management of services requests, service definition, image management, booking, instantiation of services, services discovery, monitoring, managing licenses and instance management.

The management of change and the management of configuration stores the information needed to support service management automation. This typically includes the following elements: service templates, topologies, management plans, booking and configurations.

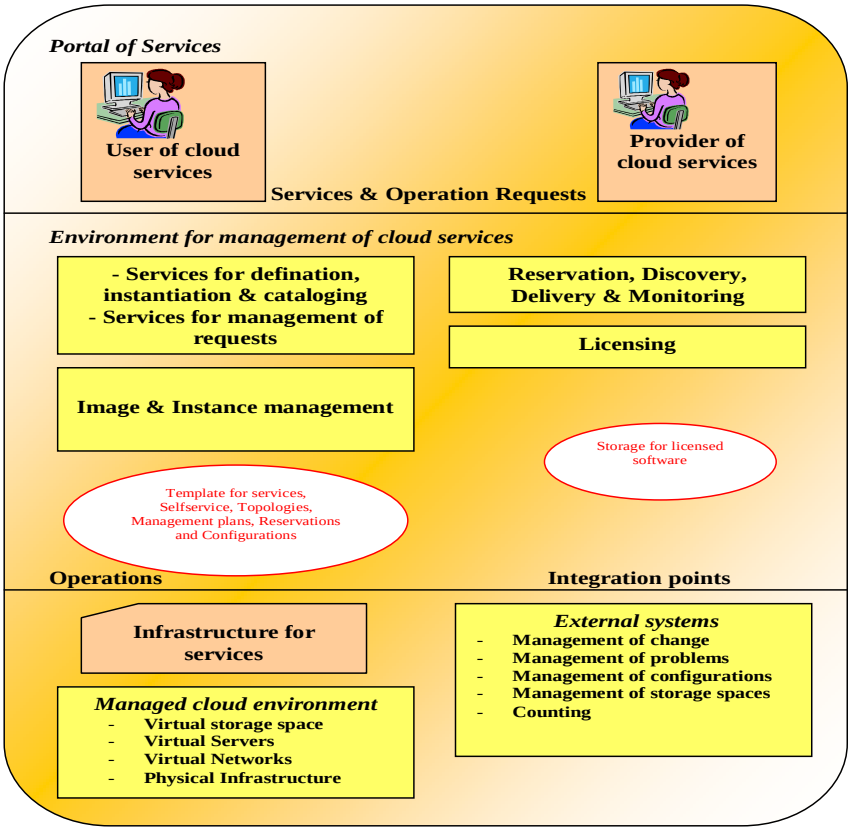

Fig. 7. The proposed private cloud architecture – the perspective of the systems involved in cloud

Managed cloud environment includes the physical and virtual hardware layer. They provide a flexible platform adapted to improve the resources utilization. In a virtualized environment, computing environments can be dynamically created, expanded, reduced, or removed depending on changes in demand. Virtualization is required for a dynamic cloud infrastructure because it provides important advantages of pooling, management and isolation of resources (multiple users and applications can share physical resources without affecting each other). Virtualization allows a set of underutilized physical servers to be consolidated into a smaller number of physical servers fully used, thereby contributing to significant cost savings. There are many forms of virtualization commonly used in today's IT infrastructures.

A common interpretation of server virtualization is the mapping of natural resources more logical representations or more partitions. The logical partitions (LPARs) and virtual machines (VM) are examples of this.

The layer of external systems represents integration points between the private cloud and the existing customer environments. Some examples of such integration points are: change management, problem

management, configuration management, storage management and metering areas.
Actions regarding the use of interoperability, open specifications and innovative e-Government were taken into account in developing the proposed architectures. Implementing collaborative solution for developing, updating and maintaining the Romanian National Interoperability Framework (RO eGIF) allows the integration of European expertise and reuse of interoperable electronic service solutions. The development of the collaborative platform and repositories involved in the development of the National Interoperability Framework was taken into account in the collaborative production of electronic services and reuse of public sector information. The open social platform contained in the proposed collaborative platform can the support action to facilitate the exchange of knowledge and experience for producing collaborative e-services and conducting activities for public sector information reuse. The private cloud architecture proposed for implementing the collaborative working environment supports the innovative e-Government action, by applying emerging technologies and paradigms.

4. Conclusions

I believe that the collaborative approach architectures proposed in this article are the best solution for Romania and for the implementation of the e-Government 2011-2015 Action Plan regarding production of the e-services by working with stakeholders and all interested parties. Proof of concept of proposed architectures may be a good starting point for the study which must be conducted in Romania according to the action plan.

Acknowledgment

This work was cofinanced from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/142115 „Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain”

References

[1] N. Benamou, A. Busson şi A. Keravel , “Impact of e-government interoperability in local governments” în Proceedings of 3rd International Conference on Electronic Government (EGOV 2004) - Electronic Government, Proceedings, 2004, pag 82-87.

[2] M. Cook şi M. LaVigne, “Making the local e-gov connection” [document online], mai 2002, [26 iunie 2006], A fost disponibil la HTTP: http://www.urbanicity.org/FullDoc.asp?ID=36

[3] F. Curbera, R. Khalaf, N. Mukhi, S. Tai & S. Weerawarana, “The next step in Web Services”, Communications of the ACM, nr. 46(10), Octombrie, pag. 29-34.

[4] e-DATA, “Extinderea studiului cu privire la stadiul electronizarii serviciilor publice”, [document online] decembrie 2005, [12 noiembrie 2007], A fost disponibil la HTTP: http://e-administratie.mcti.ro/studii/

[5] e-DATA, “Studiu cu privire la stadiul actual electronizării serviciilor în administraţia publică” [document online], iunie 2005, [12 noiembrie 2007], A fost disponibil la HTTP: http://e-administratie.mcti.ro/studii/STstadiuinf/Studiu.htm

[6] e-DATA, „Studiul pentru armonizarea procedurilor şi fluxurilor informaţionale la nivelul administraţiei publice centrale” [document online], august 2006, [12 noiembrie 2007], A fost disponibil la HTTP: http://e-administratie.mcti.ro/studii/

[7] K. D. Edmiston, “State and local e-government— Prospects and challenges”, American Review of Public Administration, nr. 33(1), Martie, 20-45.

[8] J. R. Gil-García şi L. F. Luna-Reyes, “Towards a definition of electronic

government: A comparative review" în Techno-legal aspects of the information society and new economy: An overview, A. Mendez-Vilas, J. A. Mesa Gonzalez, J. Mesa Gonzalez, V. Guerrero Bote, şi F. Zapico Alonso, Editori, Badajoz:Formatex, 2003, pag. 102-107.

[9]IDABC, "Draft for public comments – As Basis for EIF 2.0 – 15/07/2008" [document online], iulie 2008, [26 mai 2009], Disponibil la HTTP: http://ec.europa.eu/idabc/en/document/7728

[10]IDABC, "European Interoperability Framework for European public services" [document online], noiembrie 2004, [26 decembrie 2010], Disponibil la HTTP: http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf

[11]IDABC, "NIFO project" [document online], mai 2009, [26 mai 2009], Disponibil la HTTP: http://ec.europa.eu/idabc/en/document/7796

[12]K. Layne şi J. Lee, "Developing fully functional egovernment: A four stage model", Government Information Quarterly, nr. 18(2), Mai, pag. 122-136.

[13]J. Makolm, "Best practice in e-government", în Proceedings of the 1st International Conference on Electronic Government (EGOV 2002) - Electronic Government, Proceedings, 2002, pag. 370-374.

[14]J. Millard, "E-government strategies: Best practice reports from the European front line", în Proceedings of the 1st International Conference on Electronic Government (EGOV 2002) - Electronic Government, Proceedings, 2002, pag. 298-306.

[15]D. Pfaff şi B. Simon, "New services through integrated e-government", în Proceedings of the 1st International Conference on Electronic Government (EGOV 2002) - Electronic Government, Proceedings, 2002, pag. 391-394.

[16]S. H. Schelin, "E-government: An overview" în Public information technology: Policy and management issues, G. D. Garson, Ed., Hershey: Idea Group Publishing, 2003, pag. 120-137.

[17]R. Traunmueller şi M. Wimmer, "Directions in E-Government: Processes, Portals, Knowledge" în 12th International Workshop on Database and Expert Systems Applications, 2001, pag 313 – 317.

[18]UNDP, "eGov Interoperability: A review of Government interoperability frameworks in selected countries" [document online], iunie 2007, [26 mai 2008], Disponibil la HTTP: http://www.apdip.net/projects/gif/GIF-Review.pdf

[19]UNDP, "The eGovernment Interoperability Guide" [document online], iunie 2008, [26 august 2009], Disponibil la HTTP: www.apdip.net/projects/gif/GIF-Guide.pdf

Codrin-Florentin NISIOIU graduated the University "Dunarea de Jos", Galati, Faculty of Electric Engineering and Computer Science, profile - Systems and Computers Science in 2003. He got the title of doctor in Cybernetics and Economic Statistics in 2011. He had a master degree in Economic Information Systems. He is associate professor in the Economic Informatics and Cybernetics Department of the Bucharest Academy of Economic Studies. He published over 25 articles, 5 of them are included in international databases or in international catalogs. He is professional member of AIS, ACM, IEEE Computer Society and Inforec. His interests include: e-government, e-services, e-learning, e-competences and business process managemen

Stock Market Prediction using Artificial Neural Networks. Case Study of TAL1T, Nasdaq OMX Baltic Stock

Hakob GRIGORYAN
Bucharest University of Economic Studies, Bucharest, Romania
Grigoryanhakob90@yahoo.com

Predicting financial market changes is an important issue in time series analysis, receiving an increasing attention in last two decades. The combined prediction model, based on artificial neural networks (ANNs) with principal component analysis (PCA) for financial time series forecasting is presented in this work. In the modeling step, technical analysis has been conducted to select technical indicators. Then PCA approach was applied to extract the principal components from the variables for the training step. Finally, the ANN-based model called NARX was used to train the data and perform the time series forecast. TAL1T stock of Nasdaq OMX Baltic stock exchange was used as a case study. The mean square error (MSE) measure was used to evaluate the performances of proposed model. The experimental results lead to the conclusion that the proposed model can be successfully used as an alternative method to standard statistical techniques for financial time series forecasting.
Keywords: *artificial neural networks, NARX, principal component analysis, financial time series, stock prediction*

1 Introduction

Nowadays, financial time series prediction is an important subject for many financial analysts and researchers as accurate forecasting of different financial applications play a key role in investment decision making. Stock market prediction is one of the most difficult tasks of time series analysis since the financial markets are influenced by many external social-psychological and economic factors [1]. Efficient market hypothesis states that stock price movements do not follow any patterns or trends, and it is practically impossible to predict the future price movements based on the historical data [2].

However, financial time series are generally non-stationary, complicated and noisy, it is possible to design mechanisms for prediction of financial markets [3]. Technical analysis with statistical and machine learning techniques have been applied to this area in order to develop some strategies and methods to be helpful for financial time series forecasting. The statistical methods include autoregressive conditional heteroskedasticity (ARCH) [4], autoregressive integrated moving average (ARIMA) or Box-Jenkins model [5], and Smooth Transition Autoregressive (STAR) model [6]. In the area of stock prediction, feature selection plays a significant role in forecasting accuracy and efficiency. The main techniques for feature extraction include Principal Component Analysis (PCA) [7], Independent Component Analysis (ICA) [8]. Technical analysis assumes that past values of the stock have an influence on the future evolution of the market. In technical analysis, technical indicators created by special formulas are used to predict stock trends. In the past decades, complex machine learning techniques have been presented for time series prediction. Among them, artificial neural networks (ANNs) [9], Support vector machines (SVMs) [10], Genetic Algorithms [11] and Self Organizing Maps (SOM) [12] are the most common used machine learning techniques in financial time series prediction.

Since the early 1990's, ANNs have become the most popular machine learning techniques used as alternative to standard statistical models in financial time series analysis and prediction. Schoeneburg et al, [13] investigated the possibility of stock

price prediction on a short term basis by different neural networks algorithms. Their results showed that neural networks can be successfully applied to design prediction models in financial time series analysis. Kimoto et al, [14] developed a prediction system based on modular neural networks for stocks on the Tokyo Stock Exchange and showed good experimental results. Kuan et al, [15] analyzed the potential of feed-forward and recurrent neural networks in forecasting the foreign exchange rate data. Chen et al,[16] examined several neural networks to evaluate their capability in stock price and trend prediction, and concluded that class-sensitive neural network (CSNN) is the best performing neural network in both cases. D. Olson et al, [17] compared NN forecasts of one-year-ahead Canadian stock returns with the prediction results obtained using logistic regression (logit) and ordinary least squares (OLS) techniques. Their results showed that back-propagation neural networks outperform other models in classification purposes and can be used in various trading rules. M. Ghiassi et al, [18] proposed a dynamic neural network model for forecasting time series events and showed that ANN-based dynamic neural network model is more accurate and performs significantly better than the traditional ANNs and autoregressive integrated moving average (ARIMA) models.

Other authors used hybrid techniques combining ANNs with different feature extraction techniques in financial market prediction. Among them, Abraham et al, [19] used PCA as a pre-processing step for hybrid system based on neural networks and neuro-fuzzy approaches for stock market prediction and trend analysis. Aussem al, [20] proposed a combined forecast model based on wavelet transform and neural networks. They used wavelet transform to decomopose the original data into varying scales of temporal resolution and then used dynamic recurrent neural network (DRNN) to forecast S&P500 stock closing prices. Chen and Shih, [21] applied SVMs and Back Propagation (BP) neural networks to predict Asian stock market indices and showed that both models perform better than the statistical autoregressive AR models. Zhao et al, [22] proposed a wavelet neural network to forecast Shanghai stock market returns and compared their results with back propagation neural network (BP) results. They showed that the simulation result of wavelet neural network is more accurate than that of BP neural network.

More recent studies include: Lu, [23] proposed a hybrid technique with ICA and neural network model for stock price prediction. The model used ICA for denoising the time series data and the rest of ICs used to build the neural network model. Kara et al, [24] compared two models based on ANNs and SVMs in prediction of directional movements in the daily Istanbul Stock Exchange (ISE) National 100 Index and concluded that ANN model performs better than SVM model. A. Fagner et al., [25] applied a neural network based model for the short term prediction of change in direction (POCID) of closing prices of the financial market, combining technical and fundamental analysis. Wang et al, [26] used a stochastic time effective function neural network (STNN) with PCA to forecast different stock indices. Their results displayed better performance of proposed two-stage model compared with standard neural network models.

This paper presents an integrated method based on PCA and ANNs for financial time series prediction. Considering the fact that the optimal variable search plays an important role for better accuracy of forecasting results, technical analysis has been conducted to calculate technical indicators helping to predict the stock prices. The proposed approach first uses PCA technique to extract principal components from the various technical indicators then uses the filtered variables as the input of

ANN-based technique to build the forecasting model. In order to evaluate the prediction accuracy of the proposed model, the mean squared error (MSE) measure was used as an evaluation metric. The historical data set was selected from Nasdaq OMX Baltic stock exchange.

The rest of the paper is organized into five chapters: Chapter 2 introduces a dynamic neural network called nonlinear autoregressive network with exogenous input (NARX). Chapter 3 describes the research methodology, including data collection, data normalization, technical analysis, principal component analysis (PCA) and evaluation metric. Chapter 4 presents the summarized and discussed experimental results. Finally, Chapter 5 concludes the research results and presents the future work.

2. Nonlinear autoregressive network with exogenous input (NARX)

The nonlinear autoregressive network with exogenous input (NARX) is a recurrent dynamic network, with feedback connections encompassing multiple layers of the network (figure 1).

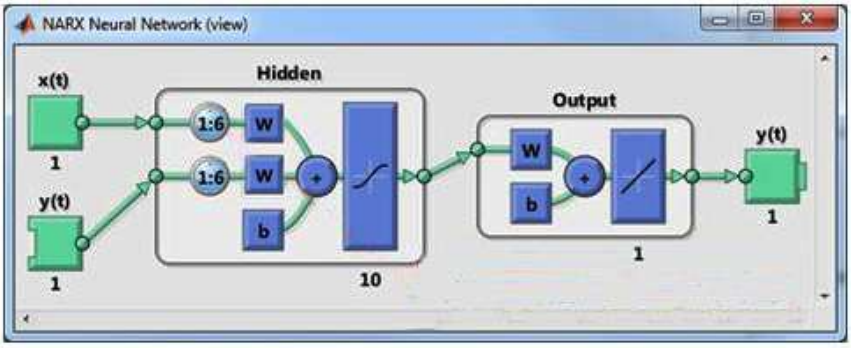

Fig. 1. The architecture of nonlinear autoregressive network with exogenous inputs (NARX)

The NARX model can be mathematically described as,

$$y(t) = f\left(y(t-1), y(t-2), \ldots, y(t-n_y), u(t-1), u(t-2), \ldots, u(t-n_u)\right) \quad (1)$$

where, *y* is the variable of interest and *u* is externally determined variable that influences the *y*. The previous values of *y* and *u* help to predict future values of *y*.

The prediction model can be defined as,

$$\hat{Y}_{(t+p)} = f_{ANN}\left(Y_t^{(d)}, X_t^{(d)}\right) \quad (2)$$

$$Y_t^{(d)} = \{Y_t, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-d}\} \quad (3)$$

$$X_t^{(d)} = \{X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-d}\} \quad (4)$$

where Y_t is the stock closing value at the moment of time t. $\hat{Y}_{(t+p)}$ is the forecasted value of the stock price for the prediction period p, and d is the delay expressing the number of pairs $(X_k, Y_k), k = t, t-1, \ldots, t-d$ used as input of the neural model. For each *t*, we denote by $X_t = (X_t(1), X_t(2), \ldots, X_t(n))^t$ the vector whose entries are the values of the indicators significantly correlated to Y_t.

In this study, the network training function is carried out by an improved backpropagation method proposed by Plagianakos et al. in [27].

3. Proposed methodology

3.1 Data collection

The research data used in this study is historical data taken from the Nasdaq OMX Baltic stock exchange and accurately chosen technical indicators. The whole data set covers the period from March 12, 2012 to December 30, 2014, a total of 700 daily observations. The historical data consists of daily closing price, opening price, lowest,

highest prices, traded volume, turnover data of Tallink Grupp AS shares (symbol TAL1T) and 30 indicators chosen from technical analysis of the stock market. Tallink stock closing price was used as a forecasting variable for this research. Historical data was collected from the Nasdaq OMX Baltic official website.

3.2 Technical analysis

Technical analysis is a security analysis method for directional prediction of prices by analyzing the historical data [28]. In other words, technical analysis relies on the assumption that past trading variables, such as price and volume can help to forecast future market trends. A technical indicator is a fundamental part of technical analysis. It presents a mathematical calculation based on the historical data.

There are in total 30 technical indicators used in this research. The complete list of all calculated technical indicators and stock based variables are given in Table 1. Some of these indicators are chosen as input variables of forecast model. The feature selection process is described in the section 3.4.

3.3 Data normalization

As the collected data has different values with different scales, it is necessary to adjust and normalize the time series at the beginning of the modelling for improving the network training step. The data normalization range is chosen to be [0,1] and the equation for data normalization is given by,

$$V = \frac{v - v_{min}}{v_{max} - v_{min}} \quad (5)$$

where V is the normalized data, v is the original data value, v_{max} and v_{min} are maximum and minimum values of the series.

Table 1. List of all 36 variables used in PCA

Technical indicator	TALL1T stock
Bollinger Bands	Opening price
Exponential Moving Average (EMA)	Closing price
Kaufman Adaptive Moving Average (KAMA)	Highest price
Simple Moving Average (MA)	Lowest price
Weighted Moving Average (WMA)	Turnover
Triangular Moving Average (TRIMA)	Traded volume
On Balance Volume (OBV)	
Average True Range (ATR)	
Average Directional Movement Index (ADX)	
Absolute Price Oscillator (APO)	
AROON	
Balance Of Power (BOP)	
Commodity Channel Index (CCI)	
Chande Momentum Oscillator (CMO)	
Directional Movement Index (DX)	
Moving Average Convergence Divergence (MACD)	
Money Flow Index (MFI)	
Momentum	
Percentage Price Oscillator (PPO)	
Rate Of Change (ROC)	
Relative Strength Index (RSI)	
%K stochastic oscillator	

%D	
Ultimate Oscillator	
Williams %R	
Minus Di	
Plus Di	
Minus Dx	
Plus Dx	
Chaykin oscillator	

3.4 Principal component analysis

The feature selection process is one of the important parts of the prediction model. It is used to filter irrelevant features from the given data set in order to to improve the prediction accuracy. Principal component analysis (PCA) is a statistical technique for feature extraction and data representation.

The main idea in PCA is to find the component vectors that explain the maximum possible amount of variance by linearly transformed components.

In signal processing, PCA can be defined as a transformation of a given set of *n* input vectors with the same length *K* formed in the *n*-dimensional vector $x = (X_1, X_2, \dots, X_n)^T$ into a vector *y* by:

$$y = A(x - \mu_x), \quad (6)$$

where the vector μ_X is the vector of the means of the input variables *x*.

The matrix A is determined by the covariance matrix C_x as the orthonormal rows of matrix A are formed from the eigenvectors of the matrix C_x.

The covariance matrix can be calculated by the equation:

$$C_x = E\left\{(x - \mu_x)(x - \mu_x)^T\right\} \quad (7)$$

Let $x = (X_1, X_2, \dots, X_n)^T$ be the n-dimensional random vector, and $a_1, a_2, \dots, a_n$ be the corresponding eigenvectors of correlation matrix *R* where the covariance between X_i and X_j is given by,

$$Cov(X_i, X_j) = \Sigma_{i,j} = \sigma^2 R_{i,j} = \lambda_j a_i^T a_j$$

, for i,j=1,2,...,n. (8)

Define W_1 to be the first principal component of the sample *x* by the linear transformation,

$$W_1 = a_1^T x = \sum_{i=1}^{n} a_{i1} x_i, \quad (9)$$

where the vector $a_1 = (a_{11}, a_{21}, \dots, a_{n1})$ and $a_1^T x = 1$.

It follows that, the first principal component W_1 has the highest possible variance $var[W_1] = a_1^T R a_1$ and the largest eigenvalue among all linear combinations of the *x*, such that $var[W_1] \geq var[W_2] \geq \cdots \geq var[W_n]$, $\lambda_1 > \lambda_2 > \dots > \lambda_n > 0$.

The problem of computing the principal components of a certain dataset can be solved many ways. Clearly, we can directly apply the above mentioned result and compute the principal components based on the correlation matrix resulted from the available data. This way, the quality of the resulted principal components depends on the distance between the theoretical correlation matrix and the one computed from data.

Some alternative strategies, for example specialized neural networks, have been proposed to perform principal component analysis (PCA) tasks. The study of the convergence properties of different stochastic learning PCA algorithms is usually performed by reducing the problem to the analysis of asymptotic stability of a dynamic system trajectories. The evolution of such systems is described in terms of an ODE. The Generalized Hebbian Algorithm (GHA) extends the Oja's learning rule for learning the first principal components using the Hotteling deflation technique. A series of experimentally established conclusions

regarding the performance and efficiency of some of the most frequently used PCA learning algorithms implemented on neural architectures are reported in [29].

3.5 Performance criteria

The prediction performance is evaluated using the mean square error (MSE) evaluation method:
MSE error measure is defined by,

$$MSE(T,P) = \frac{1}{n}\sum_{i=1}^{nr}(T(i)-P(i))^2 \quad (10)$$

where $T = (T(1), T(2), \ldots, T(n))$ is the vector of target values, $P = (P(1), P(2), \ldots, P(n))$ is the vector of predicted values and n is the number of data samples.

3.6 Prediction model based on NARX and PCA techniques

In this study, a two-stage prediction model combining PCA and NARX techniques is presented for stock market prediction. Fig 2 is the outline of the proposed prediction model. First, pre-processing step is applied to data which include: technical analysis to select proper technical indicators, data normalization to adjust and normalize the data set, and principal component analysis for features selection and data reduction. Second, the NARX model is constructed based on the feature subset from PCA. Then, data sample is trained by series-parallel architecture. After the training step, the series-parallel architecture is converted into a parallelized network, in order to execute the forecast task. The experimental results based on this model are given in the next chapter.

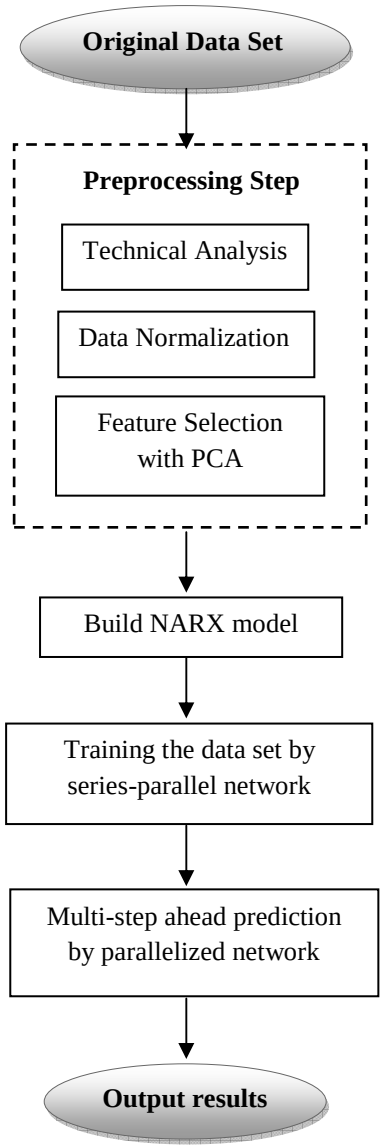

Fig. 2. The architecture of the prediction model based on NARX and PCA

4. Experimental results

Experiments have been conducted to evaluate the performance of the presented method. A data set from Nasdaq OMX Baltic stock exchange was used to conduct the experiments. The whole data set covers the period from March 12, 2012 to December 30, 2014, a total of 700 daily observations. Data set includes the traded volume, turnover, opening, closing, highest, lowest prices and 35 technical indicators. Experiments include two different forecasting periods of the same data set. The first experimental data set is divided into two parts. The first part (500 pairs of observations) is used for training, testing

and validation phases. The second part (200 pairs of observations) is reserved for prediction step. The second experiment uses 550 data samples for training and validation tests and 150 data samples for prediction task. The goal of the experiment is to predict the closing price of the Tallink stock (symbol TALL1T).

The input variables are selected by feature selection method called PCA. In PCA, each component is uncorrelated with all of the preceding components. For this reason we will have maximally uncorrelated variables used as input for prediction models. The scientific data analysis software called PAST was used to implement PCA for time series. In PCA, the correlation coefficients provide the measure of the relation between considered 36 variables. Table 2 shows the total variance of the original variables. The cumulative contribution to the total explained variance of 10 largest eigenvalues out of 36 components is 96.6% which is much higher than the normal criterion 70%. Thus, the first 10 principal components provide the most information of original data and can be selected to form the output subset.

After pre-processing step, the NARX model was developed for data training and prediction. Experiments with NARX prediction model were performed by using the software MATLAB. In the experiments with NARX model with the architecture 2×10×1, where input variables are chosen to be 10 variables which correspond to the selected 10 PCs, and the output variable is the closing price of Tallink stock. Delay was set equal to 2 using auto-correlation function of all variables. The number of neurons in the hidden layer is set according to the following equation, $2\sqrt{(m+2)N}$ where *m* stands for the number of the neurons of the output layer and *N* is the dimension of input data.

Table 2. Principle components of 36 variables

PC	*Eigenvalue*	Variance %	Cum. variance %
1	1.04985	49.796	49.796
2	0.500825	23.755	73.551
3	0.133046	6.3105	79.8615
4	0.097486	4.6239	84.4854
5	0.070471	3.3425	87.8279
6	0.05544	2.6296	90.4575
7	0.0477	2.2625	92.72
8	0.034688	1.6453	94.3653
9	0.023855	1.1315	95.4968
10	0.022894	1.0859	96.5827
11	0.016476	0.78147	97.36417
12	0.010009	0.47471	97.83888
13	0.009175	0.43518	98.27406
14	0.007363	0.34925	98.62331
15	0.006329	0.30021	98.92352
16	0.004273	0.20266	99.12618
17	0.003963	0.18796	99.31414
18	0.003799	0.18019	99.49433
19	0.003026	0.14353	99.63786
20	0.002596	0.12311	99.76097
21	0.001872	0.088811	99.84978
22	0.00115	0.054523	99.9043
23	0.000815	0.038637	99.94294
24	0.000445	0.021128	99.96407
25	0.000225	0.010659	99.97473
26	0.000143	0.0067673	99.9815
27	0.000122	0.0057873	99.98728
28	8.44E-05	0.0040046	99.99129
29	6.81E-05	0.0032319	99.99452
30	5.58E-05	0.0026458	99.99716
31	3.10E-05	0.0014708	99.99864
32	2.45E-05	0.0011619	99.9998
33	1.14E-05	0.00053899	99.99982
34	4.91E-06	0.00023269	99.99985
35	1.61E-07	7.66E-06	99.99992
36	6.23E-34	2.96E-32	100

Figure 3 and 5 show the values of the TALL1T stock and the predicted values and horizon. Figure 4 and 6 show the predicted and actual values of TALL1T stock with PCA-NARX model in two different forecast time periods. The blue line is the predictions of the proposed model and the green line is the actual values.

From the figures, it can be observed that this method forecasts values closely to the actual values in most of the time period.
The prediction performances are evaluated using the standard evaluation measure called mean squared error (MSE).

$MSE_{150} = 0.0011703$
$MSE_{200}=0.0034567$

The experimental results show that this method is effective and efficient in forecasting stock prices compared with other research studies in the field of stock market prediction.

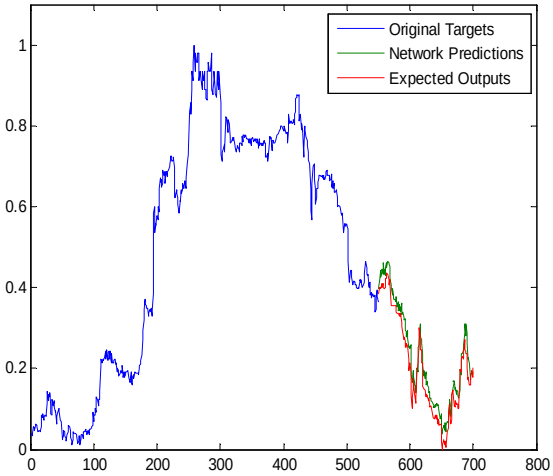

Fig. 3. Prediction results of TALL1T stock in case of 150 samples

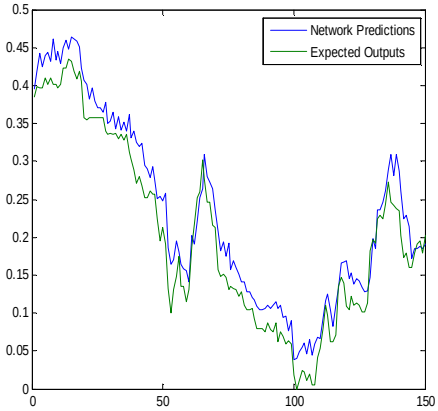

Fig. 4. Actual data and predicted data comparison of 150 samples

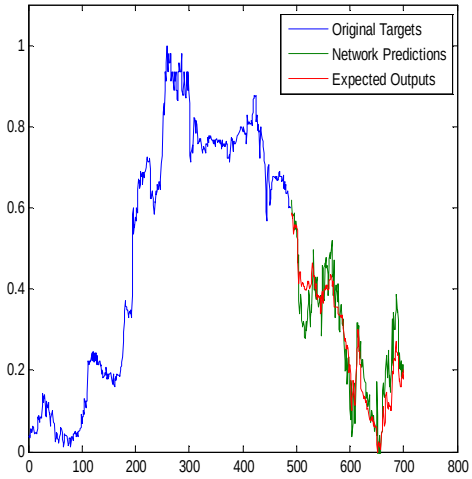

Fig. 5. Prediction results of TALL1T stock in case of 200 samples

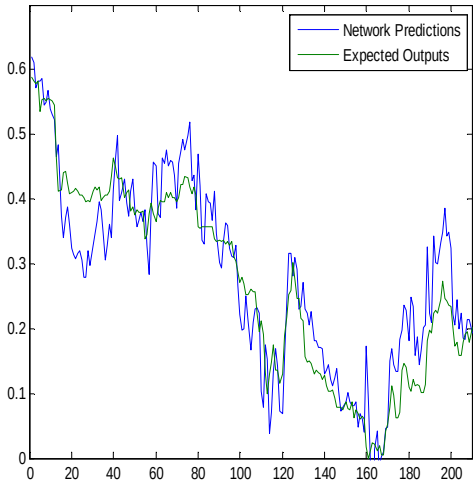

Fig. 6. Actual data and predicted data comparison of 200 samples

5. Conclusions and future work

Forecasting stock market changes is an important issue for many researchers and investors. Moreover, it is one of the challenging tasks of nowadays time series analysis. In this paper, we have presented the hybridized prediction model for financial time series forecasting. In special case study, 30 technical indicators were calculated based on technical analysis. The prediction model was constructed based on two-stage architecture, combining principal component analysis (PCA) and artificial neural networks (ANNs). This study used PCA to select proper input variables from technical indicators, and NARX model to forecast the future values in stock exchange. The experimental results obtained using the

proposed neural network approach proved better results from the point of view of MSE measure. This study allows us to conclude that PCA-NARX prediction model provides a promising alternative tool to other ANN based methods in financial time series forecasting.

This research considers only technical indicators and stock based information as stock affecting indicators. In the future work it is designed to include more stock market influencing factors, based specifically on fundamental and technical analyses compared with other prediction models.

Acknowledgment

The author would like to thank Mr. Leo Võhandu, the professor emeritus of Tallinn University of Technology for his useful suggestions and comments in statistical data analysis.

References

[1] D.N. Gujarati, Basic econometrics, McGraw-Hill, New York (2003)

[2] E.F. Fama, Efficient capital markets: A review of theory and empirical work, The Journal of Finance, 25 (2) (1970), pp. 383–417

[3] Yaser S.A.M, Atiya A.F, Introduction to financial forecasting, Applied Intelligence 1996, 6:205–13.

[4] Engle, Robert F. (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation", Econometrica 50 (4): 987–1007.

[5] G. E. P. Box and G. M. Jenkins (1976) Time Series Analysis: Forecasting and Control, Revised Edition, San Francisco: Holden Day.

[6] Teräsvirta T., (1994). Specification, estimation, and evaluation of smooth transition autoregressive models, Journal of the American Statistical Association 89, 208-218.

[7] Jolliffe, I. T. (1986), Principal Component Analysis, Springer, New York

[8] Hyvarinen A., Karhunen J., & Oja E., (2001a). Independent component analysis. New York: John Wiley and Sons.

[9] White, H. (1988). Economic prediction using neural networks: The case of IBM stock prices. Proceedings of the Second Annual IEEE Conference on Neural Networks, pp. II:451–458. New York: IEEE Press

[10] V. Vapnik. Statistical Learning Theory. John Wiley and Sons, Inc., New York, 1998.

[11] J. Holland, Adaption in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI, 1975.

[12] Kohonen, Teuvo (1982). "Self-Organized Formation of Topologically Correct Feature Maps". Biological Cybernetics 43 (1): 59–69.

[13] Schoeneburg, E., Stock Price Prediction Using Neural Networks: A Project Report, Neurocomputing, vol. 2, 1990, pp. 17-27.

[14] Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M., 1990. Stock Market prediction system with modular neural networks. In: Proceedings of the IEEE International Joint Conference on Neural Networks, San Diego, California, 2, pp. 11–16

[15] Kuan, C-M and T. Liu (1995), 'Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks', Journal of Applied Econometrics, 10(4): 347–64.

[16] C.H. Chen, Neural networks for financial market prediction, Proceedings of the IEEE International Conference on Neural Networks, 2 (1994), pp. 1199–1202

[17] D. Olson, C. Mossman, Neural network forecasts of canadian stock returns using accounting ratios, Int. J. Forecast., 19 (2003), pp. 453–465

[18] M. Ghiassi, H. Saidane, D.K. Zimbra, A dynamic artificial neural network

model for forecasting time series events, Int. J. Forecast., 21 (2005), pp. 341–362

[19] Abraham, A., B. Nath And P. K. Mahanti. (2001) "Hybrid Intelligent Systems For Stock Market Analysis," Computational Science, Springer-Verlag Germany, Vassil N. Alexandrov Et. Al. (Eds.), ISBN 3-540-42233-1, San Francisco, USA, Pp. 337-345.

[20] Aussem, A., Campbell, J., and Murtagh, F. (1998): "Wavelet-based Feature Extraction and Decomposition Strategies for Financial Forecasting," Journal of Computational Intelligence in Finance.

[21] Chen, W-H. & Shih, J.Y. (2006). Comparison of support vector machines and back propagation neural networks in forecasting the six major Asian stock markets. International Journals Electronics Finance, 1, 49-67.

[22] Y. Zhao, Y. Zhang, C. Qi, Prediction Model of Stock Market Returns Based on Wavelet-Neural Network, in: Proceedings of 2008 Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Wuhan, China, 2008, pp. 31–36.

[23] Chi-Jie Lu, Integrating independent component analysis-based denoising scheme with neural network for stock price prediction, Expert Systems with Applications 37 (2010) 7056–7064

[24] Yakup Kara, Melek Acar Boyacioglu, Ömer Kaan Baykan, Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange, Expert Systems with Applications 38 (2011) 5311–5319

[25] Fagner A. de Oliveira, Cristiane N. Nobre, Luis E. Zarate Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index – Case study of PETR4, Petrobras, Brazil, Journal of Expert Systems with Applications, 40, (2013), 7596–7606.

[26] Z. Liao, J. Wang, Forecasting model of global stock index by stochastic time effective neural network, Expert Systems with Applications, 37 (1) (2009), pp. 834–841.

[27] V.P. Plagianakos, D.G. Sotiropoulos, and M.N. Vrahatis, An Improved Backpropagation Method with Adaptive LearningRate, University of Patras, Department of Mathematics, Division of Computational Mathematics & Informatics, GR-265 00, Patras, Greece (1998).

[28] Murphy, John J. (1999). Technical analysis of the financial markets: a comprehensive guide to trading methods and applications (2nd ed.), New York institute of finance.

[29] Cătălina-Lucia Cocianu, Luminiţa State, Panayiotis Vlamos, Neural Implementation of a Class of PCA Learning Algorithms, Economic Computation and Economic Cybernetics Studies and Research, Vol. 43, No 3/2009, pp. 141-154

Hakob GRIGORYAN has graduated the Faculty of Cybernetics of the State Engineering University of Armenia (Polytechnic) in 2011. In 2014 he graduated the Faculty of Informatics and Mathematics of the University of Bucharest with a specialization in Database and Web Technologies. At the present, he is earning his PhD degree in Economic Informatics at Bucharest University of Economic Studies, coordinated by Professor Catalina-Lucia Cocianu. His PhD thesis is "Machine Learning-Based Techniques for Financial Data Analysis and Forecasting Purposes".

Enhancing Forecasting Performance of Naïve-Bayes Classifiers with Discretization Techniques

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

During recent years, the amounts of data, collected and stored by organizations on a daily basis, have been growing constantly. These large volumes of data need to be analyzed, so organizations need innovative new solutions for extracting the significant information from these data. Such solutions are provided by data mining techniques, which apply advanced data analysis methods for discovering meaningful patterns within the raw data. In order to apply these techniques, such as Naïve-Bayes classifier, data needs to be preprocessed and transformed, to increase the accuracy and efficiency of the algorithms and obtain the best results.

This paper focuses on performing a comparative analysis of the forecasting performance obtained with the Naïve-Bayes classifier on a dataset, by applying different data discretization methods opposed to running the algorithms on the initial dataset.

Keywords: *Discretization, Naïve-Bayes classifier, Data mining, Performance*

1 Introduction

Nowadays, organizations collect large amounts of data every day. These data need to be analyzed in order to find the meaningful information contained by it and reach the best conclusions, to support decision making.

Data mining is an innovative new solution, which provides the required tools for processing the data in order to extract significant patterns and trends. Before running data mining algorithms against it, the raw data needs to be cleaned and transformed. This is accomplished through the preliminary steps of the Knowledge Discovery in Databases process – data preprocessing and data transformation. One of the key methods used during data transformation is data discretization.

Discretization methods transform the continuous values of a dataset attribute to discrete ones. It can help improve significantly the forecasting performance of classification algorithms, like Naïve Bayes, that are sensitive to the dimensionality of the data.

Naïve-Bayes is an intuitive data mining algorithm that predicts class membership, using the probabilities of each attribute value to belong to each class.

Discretization methods needs to be applied on datasets before analyzing them, in order to transform the continuous variables to discrete variables and, thus, to improve the accuracy and efficiency of the classification algorithm.

2. Naïve-Bayes classifiers: overview

Classification is a fundamental issue in machine learning and statistics. It is a supervised data mining technique, with the goal of accurately predicting the class label for each item in a given dataset. A classification model built to predict class labels, from the attributes of the dataset, is known as a classifier.

In data mining, Bayesian classifiers are a family of probabilistic classifiers, based on applying Bayes' theorem. The theorem, named after Reverend Thomas Bayes (1701–1761), who has greatly contributed to the field probability and statistics, is a mathematical formula used for calculating conditional probabilities. It relates current probability to prior probability.

Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Studies comparing

classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. [1]

The Naïve-Bayes classifier is an intuitive data mining method that uses the probabilities of each attribute value belonging to each class to predict class membership.

A Bayesian classifier is stated mathematically as the following equation: [1]

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (1)$$

where,

- $P(C_i|X)$ is the probability of dataset item X belonging to class C_i;
- $P(X|C_i)$ is the probability of generating dataset item X given class C_i;
- $P(C_i)$ is the probability of occurrence of class C_i;
- $P(X)$ is the probability of occurrence of dataset item X.

Naïve-Bayes classifiers simplify the computation of probabilities by assuming that the probability of each attribute value to belong to a given class label is independent of all the other attribute values.

This method goes by the name of Naïve Bayes because it's based on Bayes' rule and "naïvely" assumes independence—it is only valid to multiply probabilities when the events are independent. The assumption that attributes are independent (given the class) in real life certainly is a simplistic one. But despite the disparaging name, Naïve Bayes works very effectively when tested on actual datasets, particularly when combined with some of the attribute selection procedures. [2]

3 Discretization techniques: theoretical framework

The Knowledge Discovery in Databases (KDD) process is an iterative process for identifying valid, new and significant patterns in large and complex datasets. The core step of the KDD process is data mining, which involves developing the model for discovering patterns and trends in the data.

Data preprocessing and data transformation are crucial steps of the KDD process. After performing them better data should be generated, in a form suitable for the data mining algorithms.

Data transformation methods include dimension reduction (such as feature selection and extraction, and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation). [3]

Most experimental datasets have attributes with continuous values. However, data mining techniques often need that the attributes describing the data are discrete, so the discretization of the continuous attributes before applying the algorithms is important for producing better data mining models.

The goal of discretization is to reduce the number of values a continuous attribute assumes by grouping them into a number, n, of intervals (bins). [4]

Mainly there are two tasks of discretization. The first task is to find the number of discrete intervals. Only a few discretization algorithms perform this; often, the user must specify the number of intervals or provide a heuristic rule. The second task is to find the width, or the boundaries, of the intervals given the range of values of a continuous attribute. [5]

Data discretization comprises a large variety of methods. They can be classified based on how the discretization is performed into: supervised vs. unsupervised, global vs. local, static vs. dynamic, parametric vs. non-parametric, hierarchical vs. non-hierarchical etc.

There are several methods that can be used for data discretization. Supervised discretization methods can be divided into

error-based, entropy-based or statistics based. Among the unsupervised discretization methods there are the ones like equal-width and equal-frequency.

- *Equal-width discretization*

This method consists of sorting the values of the dataset and dividing them into intervals (bins) of equal range. The user specifies k, the number of intervals to be calculated, then the algorithm determined the minimum and maximum values and divides the dataset into k intervals.

Equal-width interval discretization is a simplest discretization method that divides the range of observed values for a feature into k equal sized bins, where k is a parameter provided by the user. The process involves sorting the observed values of a continuous feature and finding the minimum, V_{min} and maximum, V_{max}, values. [5]

The method divides the dataset into k intervals of equal size. The width of each interval is calculated using the following formula:

$$Width = \frac{V_{max} - V_{min}}{k} \qquad (2)$$

The boundaries of the intervals are calculated as: V_{min}, V_{min}+Width, V_{min}+2Width, ... , V_{min}+(k-1)Width, V_{max}.

The limitations of this method are given by the uneven distribution of the data points: some intervals may contain much more data points than other. [5]

- *Equal-frequency discretization*

This method is based on dividing the dataset into intervals containing the same number of items. Partitioning of data is based on allocating the same number of instances to each bin. The user supplies k, the number of intervals to be calculated, then the algorithm divides n, the total number of items belonging to the dataset, by k.

Equal-Frequency Discretization predefines k, the number of intervals. It then divides the sorted values into k intervals so that each interval contains approximately the same number of training instances. Suppose there are n training instances, each interval then contains n/k training instances with adjacent (possibly identical) values. [3]

The method divides the dataset into k intervals with equal number of instances. The intervals can be computed using the following formula:

$$Interval = \frac{n}{k} \qquad (3)$$

Equal-frequency binning can yield excellent results, at least in conjunction with the Naïve Bayes learning scheme, when the number of bins is chosen in a data-dependent fashion by setting it to the square root of the number of instances. [2]

- *Entropy-based discretization*

One of the supervised discretization methods, introduced by Fayyad and Irani, is called the entropy-based discretization. An entropy-based method will use the class information entropy of candidate partitions to select boundaries for discretization. [5]

The method calculates the entropy based on the class labels and finds the best split-points, so that most of the values in an interval fit the same class label – the split-points with the maximal information gain.

The entropy function for a given set S is calculated using the formula: [5]

$$Info(S) = -\sum p_i \log_2 (p_i) \qquad (4)$$

Based on this entropy measure, the discretization algorithm can find potential split-points within the existing range of continuous values. The split-point with the lowest entropy is chosen to split the range into two intervals, and the binary split is continued with each part until a stopping criterion is satisfied. [5]

Many other discretization methods may be applied on raw data, both supervised and unsupervised. Among the supervised methods we can mention Chi-Square based discretization, while a sophisticated

unsupervised method is k-means discretization, based on clustering analysis.
Discretization techniques are generally considered to improve the forecasting performance of data mining techniques, particularly classification algorithms like Naïve-Bayes classifier, and, at the same time, it is thought that, choosing one discretization algorithm over another, influences the significance of the forecasting improvement.

4. Case study: Evaluating the performance of Naïve-Bayes classifiers on discretized datasets

This case study focuses on presenting the experimental results obtained by forecasting the class label for the Credit Approval dataset, using the Naïve-Bayes classifier.
The algorithm was applied to the original data, as well as to each transformed dataset, obtained by using each of the discretization methods described in this paper.
The dataset used for the experimental study concerns credit card applications and it was obtained from UCI Machine Learning Repository [6].
The Credit Approval dataset comprises 690 instances, characterized by 15 attributes and a class attribute. The values of the class attribute in the dataset can be "+" (positive) or "-" (negative) and they indicate the credit card application status for each submitted application.
This experimental study was performed using RapidMiner Software [7]. RapidMiner is a software platform, developed by the company of the same name, which provides support for all steps of the data mining process.
RapidMiner Software supports data discretization through its discretization operators. Five discretization methods are provided by RapidMiner: Discretize by Binning, Discretize by Frequency, Discretize by Size, Discretize by Entropy and Discretize by User Specification.
Among these, three methods were used during the case study, corresponding to the ones described in the paper: [7]

- *Discretize by Binning* – this operator discretizes the selected numerical attributes into user-specified number of bins. Bins of equal range are automatically generated, the number of the values in different bins may vary;
- *Discretize by Frequency* – this operator converts the selected numerical attributes into nominal attributes by discretizing the numerical attribute into a user-specified number of bins. Bins of equal frequency are automatically generated, the range of different bins may vary;
- *Discretize by Entropy* – this operator converts the selected numerical attributes into nominal attributes. The boundaries of the bins are chosen so that the entropy is minimized in the induced partitions.

During the experiment I defined a data mining process in RapidMiner. This process applied Naïve-Bayes classifier on the Credit Approval dataset, first on the original data and afterwards on the datasets obtained by applying each discretization method. I obtained performance indicators for each of these cases and I performed a comparative analysis on the results achieved without discretization and the results achieved with each discretization method.
The process flow defined in RapidMiner, for applying the Naïve-Bayes classifier, is presented in figure 1:

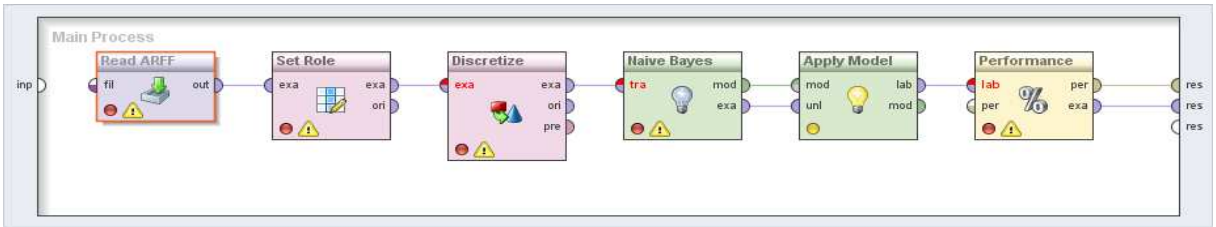

Fig. 1. Naïve-Bayes process flow in RapidMiner

The process defined for evaluating the performance of Naïve-Bayes classifier includes the following RapidMiner operators:

- *Read ARFF* – this operator is used for reading an ARFF file, in our case the file credit-a.arff;
- *Set Role* – this operator is used to change the role of one or more attributes;
- *Discretize* – this operator discretizes the selected numerical attributes to nominal attributes. RapidMiner supports applying all three discretization methods described in this paper: for equal-width discretization we can use Discretize by Binning, for equal-frequency discretization we can use Discretize by Frequency and for entropy-based discretization we can use Discretize by Entropy.
- *Naïve-Bayes* – this operator generates a Naive Bayes classification model. A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "'independent feature model". In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature [7];
- *Apply Model* – this operator applies an already learnt or trained model, in this case Naïve-Bayes, on a dataset, for prediction purposes;
- *Performance* – this operator is used for performance evaluation, through a number of statistical indicators. For my case study I chose to use accuracy, kappa and root mean squared error (RMSE).

The process described above was executed first without the Discretize operator, on the original Credit Approval dataset, and afterwards including the Discretize operator, for each discretization method. By executing the results, in each case, the process generated results for evaluating the performance of the Naïve-Bayes classifier on the dataset. Discretization methods should increase the efficiency of Naïve-Bayes classifiers. For performance evaluation I compared the results obtained in terms of accuracy, kappa and root mean square error. In order to establish which discretization technique was better for transforming the dataset, the accuracy of the classification should be as higher, as well as kappa, while the root mean square error should be as low as possible. A comparative analysis of the performance achieved, for the Naïve-Bayes classifier, with each discretization method, is shown in table 1:

Table 1. Comparative analysis of the Naïve-Bayes classifier performance with discretization methods

Discretization method	**Performance indicator**		
	Accuracy	*Kappa*	*Root mean squared error*
None	78.26%	0.546	0.432
Equal-width discretization	87.83%	0.754	0.314
Equal-frequency discretization	84.06%	0.677	0.346
Entropy-based discretization	86.96%	0.733	0.325

Based on the results in the table above, it is obvious that applying discretization methods to the dataset, before running the Naïve-Bayes classifier, has significantly improved the performance of the classification algorithm and increased the accuracy of the results obtained.

Among the discretization methods applied as part of the experimental study, equal-width discretization produces the lowest root mean squared error, 0.314, compared to the other two methods. Equal-frequency discretization generates the highest root mean squared error, of 0.346.

The prediction accuracy of the Naïve-Bayes classifier has the highest value when applying equal-width discretization - 87.83%, while equal-frequency discretization has an accuracy of 84.06%.

Performance indicator kappa compares the observed accuracy with the expected accuracy of the classifier. Thus, the best discretization method is the one generating the highest kappa. The discretization method producing the highest kappa, of 0.754, is equal-width discretization. Applying equal-frequency discretization generates the lowest kappa, 0.677.

My experiment compares the quality of the classification achieved through the Naïve-Bayes classifier, on the Credit Approval dataset, without performing discretization on the raw data and after applying discretization methods on the data. Through this experiment I am also comparing the performance obtained by applying each of the discretization methods, against each other.

Based on the previous statements, all three discretization methods improve the quality of the classification, compared to running the classification on the initial dataset. Among the methods, the best quality classification is obtained by applying equal-width discretization, opposed to equal-frequency discretization, which generates the lowest quality classification.

5. Conclusions

Discretization is a very important transformation step for data mining algorithms that can only handle discrete data. The results of my tests confirm that the performance of Naïve-Bayes classifier is improved when discretization methods are applied on the dataset used in the analysis. In this paper, I have studied the effect that applying different discretization methods, can have on the results obtained by performing a classification analysis, with the Naïve-Bayes classifier. The conclusion, based on the results obtained, is that applying the discretization methods prior to running the classification algorithm is beneficial for the analysis, since better performance indicators have been obtained on the discrete data.

Based on the experimental results, I can assert that Naïve-Bayes classifier generates better results on discrete datasets and, also, that for the particular dataset we used, the most efficient discretization method was equal-width discretization, while equal-frequency discretization was the least efficient for improving the classification efficiency and accuracy of the Naïve-Bayes classifier.

References

[1]Jiawei Han, Micheline Kamber and Jian Pei – "Data Mining: Concepts and Techniques. Third Edition", Morgan Kaufmann Publishers, USA, 2011, ISBN 978-0-12-381479-1.

[2]Ian H. Witten, Eibe Frank and Mark A. Hall - "Data Mining: Practical Machine Learning Tools and Techniques. Third Edition", Morgan Kaufmann Publishers, USA, 2011, ISBN 978-0-12-374856-0.

[3]Oded Maimon and Lior Rokach – "Data Mining and Knowledge Discovery Handbook. Second Edition", Springer Publisher, London, 2010, ISBN 978-0-387-09822-4.

[4]Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski and Lukasz A. Kurgan – "Data Mining. A Knowledge Discovery Approach", Springer

Publisher, USA, 2007, ISBN 978-0-387-33333-5. [5]Rajashree Dash, Rajib Lochan Paramguru and Rasmita Dash – "Comparative Analysis of Supervised and Unsupervised Discretization Techniques", International Journal of Advances in Science and Technology, Vol. 2, No. 3, 2011, ISSN 2229-5216.

[6]UCI Machine Learning Repository – Credit Approval Data Set, Available online (May 16th, 2015): https://archive.ics.uci.edu/ml/datasets/Credit+Approval.

[7]RapidMiner Software, Available online (May 14th, 2015): https://rapidminer.com/.

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. Her scientific fields of interest include: Databases, Data Warehouses, Business Intelligence, Decision Support Systems and Data Mining.

Big Data, indispensable today

Radu – Ioan ENACHE, Marian Adrian ENE
Academy of Economic Studies, Bucharest, Romania
radu.enache91@gmail.com, ene.marianadrian@outlook.com

Big data is and will be used more in the future as a tool for everything that happens both online and offline. Of course , online is a real hobbit, Big Data is found in this medium , offering many advantages , being a real help for all consumers. In this paper we talked about Big Data as being a plus in developing new applications, by gathering useful information about the users and their behaviour.We've also presented the key aspects of real-time monitoring and the architecture principles of this technology. The most important benefit brought to this paper is presented in the cloud section.
Keywords: *Big Data, Cloud Data, Financial Data, Data encryption*

1 About Big Data

From the beginning, data was always highly structured. All the data was divided into fields, those fields had a specific length, and the data that entered in every field was constrained by a set of rules. Today, most of the data entered by humans is unstructured, having the form of free text. This text comes email messages, tweets, documents and so on. [3]
Big Data is the latest trend emerged in data management. It is defined as any data source which has three features:

- Very large data volume;
- Very high data transfer rate;
- Huge variety of data.

Big data is very important because it enables organizations to collect, store, manage and manipulate large amounts of data at optimal speed, to acquire necessary information. This is not a single technology, but rather a combination of 50 years of technological evolution. Each wave in data management is born from the need to solve some problems like this. [1]
Stages of development for data management in the last 50 years have culminated to the point where we are today: the beginning of the big data age.

1. Creating manageable data structure

As computers have moved on the commercial market, in the late 60s, the data was stored in files that had no structure. When companies understood their consumers at a more detailed level, they had to apply brute force, including detailed models of programming to reach the required data.
Later, after the mid-70s, things have changed with the advent of the relational database management system and the relational database, requiring a structure and a method for improving performances. Finally, a relational database got to store objects of type BLOB (binary large objects).

2. Content Management and the Web

Most of the data available today aren't structured. Paradoxically, companies have turned their investments to structured data systems. Content management system,in the business field, has evolved in the '80s, so companies could have the ability to better manage unstructured data, mostly being documents.
In the '90s, with the rise of the Web, organizations wanted to move beyond documents to store and manage Web content, images, audio and video. The market has evolved from a set of disconnected solutions on a unified model that brought together these elements, in a platform that incorporates business process management. [1]

3. Administrating big data

Big data is a new technology that was derived from data management history. It is built on the evolution of 50 years of data management practices. With big data, it

can be made a big virtualization of data that can be stored efficiently, and using cloud storage methods, it can do so at a lower cost.

Variety is one of the main principles of big data. Although data management exists form a long time, in the world big data, two factors have recently emerged:

- ✓ Some big data sources are new, such as data generated by sensors, mobile or tablets;
- ✓ Data that was created in the past, was not captured or stored and analyzed in a useful manner. The main reason for this is that the technology did not exist to do this. In other words, there is an effective way in terms of the cost to analyze all that data.

The term "structured data" generally refers to data that have a length and a well-defined format. Examples of structured data can be numbers and groups of words called strings. Most experts say that this type of data represents about 20% of all existing ones. Structured data is typically stored in a database and can be queried using programming languages such as SQL. [1]

The data sources are divided into two categories:

- ➢ The ones generated by computers or machines: machines generated data refer to those created without human intervention;
- ➢ The ones generated by humans: These types of data came from the interaction of people with PCs.

Some experts argue that there is a third category, which is a hybrid between the two above.

2. Automatic generated data

The data generated by computers or machines may include the following:

- ❖ The data collected by sensors: Examples are radio waves, medical devices or GPS;
- ❖ Captured data from the web: When servers, applications and networks operate on the internet, they capture a variety of data, following their activity. Their volume can be very high and useful for creating service level agreements or to predict future security breaches;
- ❖ Information about sales: When the cashier scans the barcode of a product purchased, that product and all of it's related data are generated and recorded.
- ❖ Financial Data: Many systems are now scheduled. These are made based on predefined rules that automate processes.

Structured data that can be generated by man are:

- ❖ Input: These can include any type of data that a person can enter into a computer, such as names, ages, incomes, responses accumulated after some polls. These data can be used to understand consumer behavior.
- ❖ Data generated from clicks: These data is generated every time someone clicks on a link, within a certain website. These data can be analyzed to understand the behavior of consumers and their habits of purchase.
- ❖ The data generated from games: Every move you make in a game is recorded. These data can help understand the behavior of players in a certain application of gaming.

Real-time aspects of the big data can be revolutionary when companies need to resolve issues of significant difficulty. In general, this approach is only relevant in real time when the response to a question is urgent. It may be related to something important like a hospital detection equipment performance or anticipation of a security breach. Below is a list of some examples when some companies can benefit from this data in real time:

- ✓ Monitoring to identify some problems with certain information such as fraud and intelligence;
- ✓ News monitoring and social networks to identify events that could have a negative impact on financial markets

such as consumer reaction when a new product is announced;
- ✓ Change of location where ads are placed during a sporting event;
- ✓ Offering a discount coupon for a buyer, during a sale. [1]

In terms of data architecture, principles of good design are critical when creating or when migrating an environment to support big data, speaking of storage, reporting, analysis or applications. In the process of creating the environment, it should be considered the hardware, software infrastructure, well-defined APIs and even development programs. The architecture must be able to address all fundamental requirements:

- Capturing;
- Integrating
- Organizing;
- Analyzing;
- Action.

3. Technology layers

Big Data technology layers are:

A. Physical Infrastructure

At the lowest level of the architecture of a system, big data's physical architecture is being represented by hardware, networks and others. Big data implementations have specific requirements regarding the architecture's elements. It is important to bear in mind several principles that can be applied in this case, such as:

- Performance: This is also known as latency and is often measured with a single transaction or a single demand;
- Validity: This is a percentage and is computed based on service availability in a given period of time;
- Scalability: This is represented by the size of the infrastructure, storage space and processing power;
- Flexibility: This is represented by the time when there can be added new infrastructure resources, or how long the service can return to normal in the event of a system failure;
- Cost: This principle is laid down by the cost of equipment.

B. Infrastructure security

Privacy and security requirements of big data are similar to those of conventional media data. Security was aligned with business requirements. Sometimes unique challenges arise when the big data is part of a strategy, such as:

- Access to data: User access to the processed big data or not, has almost the same technical requirements as implementations which do not relate big data;
- Access to application: Most APIs provide protection against the use or access. The level of protection is probably adequate for most big data implementations.
- Data encryption: Data encryption is the biggest challenge related to the security of a big data system in one environment. In traditional media, encrypting and decrypting data require a lot of resources, but because of the volume, variety and velocity associated with big data technology, this issue is no longer valid.
- Threat Detection: The inclusion of mobile devices and social networks has grown exponentially the amount of data, but also the number of threats.

C. Operational databases

At the base of each environment more big data system databases contain data of all collections relevant to a particular business. These systems must be fast, scalable and resilient. These systems are not created the same, so one from an environment may be different from another in other environment. SQL is the most used programming language for querying databases, but other languages can

provide solutions to some of the big data challenges. Also, you can use other alternatives such as Python or Java. It is very important to understand what types of data are handled by the database or if it supports transactional behavior. Designers describe this database by the acronym ACID. This means:

- Atomicity: A transaction represents "all or nothing" when it is atomic. If any part of the transaction fails, the entire transaction fails;
- Consistency: Only transactions that contain valid data will run on the database. If the data is in a damaged condition, the transaction will not be completed and no data will be placed in database;
- Isolation: multiple transactions, simultaneously, will not interfere with one another. A valid transaction will be executed until it is completed, and their order of execution is influenced by the time they were sent;
- Durability: After transactional data has entered in the database, it stays there forever.

D. The organization of data and the instruments

The organization of data and tools for their capture, validation and assembly of big data components in a certain context is a relevant collection. Because big data is huge, techniques have evolved to process data more efficiently. One of these techniques is called MapReduce and is one of the most used. Organizing data services is in reality an ecosystem of tools and technology that captures and records data for further processing. Technologies in this category include the following:

- A distributed file system: It is necessary to provide scalability and storage;
- Serialization services: This is necessary for persistent storage systems and procedures to run some from a distance;
- Coordination of services: It is necessary for building distributed applications;
- Tools for extract, transform and load (ETL): They are necessary for loading structured and unstructured data in Hadoop. With YARN, Hadoop V2's Job Tracker has been split into a master Resource Manager and slave-based,ApplicationMaster processes. It separates the major tasks of the Job Tracker:resource management and monitoring/scheduling. The Job History server now has the function of providing information about completed jobs. [2]

E. Deposits of analytical data

Data warehouse and data mart are the two techniques that organizations use to optimize your time and help in making decisions. Because many data warehouse and data mart sites contain data accumulated from different sources within a tax refund companies, costs of data normalization are not ignored. With big data, there are some differences:

- Traditional data sources can produce detailed separate data;
- Lots of data sources exist, each needing a certain degree of manipulation before the data accumulated can be used in a business;
- Content sources also should be cleaned and they may need different techniques that are used in the structured data.

Existing tools and techniques for analyzing Big Data are invaluable. Since these algorithms do not work in normal parameters because of large data flows, they should be optimized.

Big data custom applications offer an alternative distribution and examination of data sources. Since all components are

important in a big data system, this is the place where it is the most activity in terms of innovation and creativity. These applications are aligned horizontally, that addresses common problems in companies or vertically oriented, intended to solve one problem. [1]

Virtualization in a big data system is very important. The separate service resources, allow the creation of several virtual systems in a single physical system. One of the reasons why companies have implemented virtualization is to improve the performance and efficiency of processing.

Using a set of distributed resources, such as servers, in a more flexible and efficient it delivers significant benefits in terms of reducing costs and increasing productivity. This practice has several benefits and among these are:

- Virtualization can improve utilization of physical resources;
- Virtualization allows better control over the use and performance of IT resources;
- Virtualization can provide a level of automation and standardization to optimize the computing environment;
- Virtualization provides the foundation for a cloud computing system.

4. Cloud and Big Data

We all know the power of cloud is that users can access whenever required storage resources with little or no IT support or the need to purchase more hardware or software . One of the key characteristics of cloud is elastic scalability : Users can add or subtract resources almost in real time based on changing requirements . Cloud plays an important role in the world of Big Data. Dramatic changes happen when these infrastructure components are combined with advances in data management .

Bed and optimized infrastructure supports the implementation of Big Data.

Cloud computing is a method of providing a set of shared computing resources including applications, computing, storage, networking, development, and deployment platforms and business processes.

A popular example of the benefits of cloud is so big that support data can be observed both Google and Amazon.com. Both companies depend on the capacity to manage massive amounts of data in bulk to make things right direction. These providers need to come up with the infrastructure and technology that could support applications so massive a scale. Consider the millions and millions of Gmail messages that Google processes daily as part of this service. Google has been able to optimize the Linux operating system and software environment to support e-mail in the most efficient way; therefore, it can easily support hundreds of millions of users.

Even more important, Google is able to capture the massive amount of data about both: users of mail and search engine users to develop business.

Two key cloud patterns are important in the discussion of Big Data - public cloud and private cloud. For organizations that adopt and implement cloud delivery models, most will use a combination of private calculation made by an external company for the sharing of a variety of customers who pay a per-use fee. How these companies argue in the public and private balance depends on a number of hatches, including privacy, latency, and purpose.

Public Cloud

Public Cloud is a set of hardware, networking, storage, services, applications, and interfaces owned and operated by a third party for use by other companies and individuals.

These commercial providers create extreme scalability of data center infrastructure that hide underlying details of the consumer.

Public cloud is viable because typically manage workloads relative or simple repetitive. For example, Email is a very simple application.

Therefore, a cloud provider can optimize the environment so as to be best suited to

support a large number of clients, even if many messages saved
In contrast, the data center (data center) to bear so many different applications and workloads that cannot be easily optimized. A public cloud can be very effective when an organization runs a complex data analysis and needs more computing cycles to handle the load. In addition, companies can choose to store data in a public cloud where the cost per gigabyte is relatively inexpensive compared with an acquiring the storage.

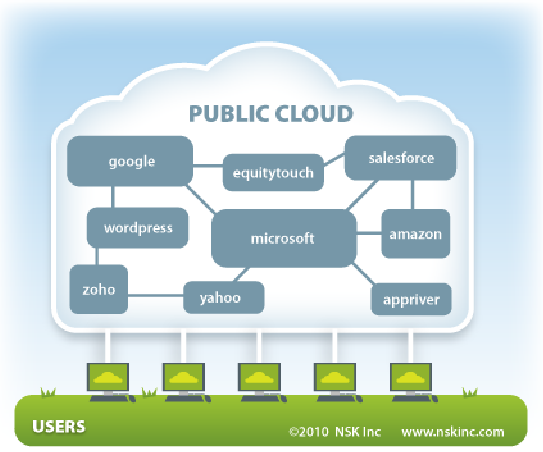

Fig. 1. Public Cloud

Therefore, all Public cloud environments are not the same. Some are scalable managed services with high security and high level service management. Other are less robust and less secure, but they are much less expensive to use.
Your choice will depend on the nature of big data projects and the amount of risk it can be assumed.

Private cloud

A private cloud is a set of hardware, network, storage, services, application and interfaces owned and operated by an organization for use by employees, partners and customers. A private cloud can be created and managed by a third party for the exclusive use of a company.
It is a highly controlled environment, not open for public consumption. The private cloud behind the firewall. Iit is highly automated with a focus on governance, security and compliance. Automation replaced manual processes of IT service management to customer support. In this way, business rules and processes can be implemented internal software so that the environment becomes more predictable and manageable.
If organizations are focused on managing a project of Big Data and applications that are processing massive amounts of data, private cloud might be the best choice in terms of latency and security.

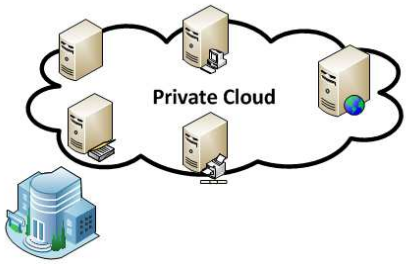

Fig. 2. Private Cloud

A Hybrid Cloud is a combination of a private cloud combined with the use of public cloud services with one or more touch points between environments .

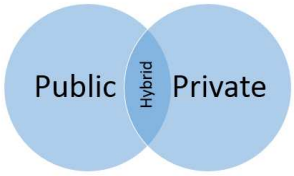

Fig. 3. Hybrid Cloud

The goal is to create a well-managed cloud environment that can combine services and data from a variety of cloud models to create a unified, automated, and well-managed computing environment.
Web companies are among the early adopters of big data, largely because of the volume of unstructured information that they must deal with on a regular basis. However, even traditional industries such as telecommunications, retail, financial services, and healthcare are launching pilots and testing the waters to see what big data has to offer. [5]
In fact, a large number of cloud features, define an ecosystem of Big Data. Here are some of them:

Scalability

Scalability in terms of hardware refers to the ability to go from small to large amounts of processing power with the same architecture. In terms of software, it refers to the consistency of performance per unit of energy and increase hardware resources. Cloud can scale to large volumes of data. Distributed computing system, an integral part of the cloud, really working on a plan to "divide and conquer". So if you have large volumes of data, they may be divided on "cloud servers".

Elasticity

Elasticity refers to the ability to expand or shrink computing resources in real time based on need. One advantage is that cloud community, customers have the potential to access a service as much as they need when they need it. This can be useful for Big Data projects which could expand the value of the resources to cope with the volume and speed data.

Of course, this feature is very attractive to end customers means that a service provider needs to design a platform architecture that is optimized for this type of service.

Pooling resources

Cloud architecture enables efficient creation of shared resource groups that make the cloud economically viable.

Pay as You Go:

A typical billing option for a cloud provider is Pay As You Go (PAYG), which means you are charged for the resources used by the court based pricing. This can be useful if you are unsure of what resources you need for your Big Data project (unless you are under budget).

Market players Cloud

Cloud Players come in all shapes and sizes to provide more differentiated products. Some are household names while others are newly emerging. Providers offering cloud services for Big Data projects are:

Amazon.com, AT & T, GoGrid, Joyent, Rackspace, IBM, and Verizon / Terremark. However, companies cloud and cloud service providers are also suppliers of software solutions specifically targeting Big Data projects.

5. Practices Big Data

While we are in an early stage in the evolution of big data, it is never too early to start with good practice, so you can set up what is learning and experience is gained. As every important emerging technology, it is important to understand why you need to leverage technology and a concrete plan.

A. Understand your goals

Many organizations begin their journey by experimenting Big date with one project that could provide some tangible benefits. By selecting a project, test freedom without risking capital expenditure. However, if you get to do a number of specific projects, you probably will not have a good plan when you begin to understand the value of leveraging Big Data within the company. Therefore, after ending some experiments and have a good initial understanding of what might be possible, you need to set some goals - both short and long term. What you want to achieve with the Big Data? It is important to have collaboration between IT and business units to better define your goals.[1]

B. Establish a trajectory

After setting goals, Amazon is a way you could define later. The company expert in exploiting analytical data to create customer intimacy as a competitive advantage. Recommendations are built on the data type "Might you interesting and ..." "Those who bought this item also bought ...". It is an example of using exploited getting better and ecommerce players in Romania.

C. Discover your data

No company ever complains you hold too little data. In reality, swim in date. The problem is that some companies do not know to use these data to predict future pragmatic, execute important business processes, or simply gain new insights.

Big time strategy aim must be to find a way to leverage data for business results more predictable. But it must go forward,

and start by embarking on a process of discovery. This process will provide a lot of perspectives.

For example, let you know how many sources of data you have and how much overlap there. This process will also help you understand who those sources goals.

D. Understanding of technological options

Now, you understand your company's objectives, have an understanding of what data you have, and you know what data are missing. But as take steps to execute the strategy? You must know what technologies are available and how they could help the company produce better results. Therefore, do your homework.

Begin to understand the value of technologies like Hadoop, offers products and complex data processing streaming events. You should look at the different types of databases, such as memory databases, spatial databases, and so on. You should familiarize yourself with the tools and techniques that are being developed as part of the big data ecosystem. It is important that your team has an understanding of the technology available to make informed choices.

Public health

The use of big data can improve public health surveillance and response. By using a nationwide patient and treatment database, public health officials can ensure the rapid, coordinated detection of infectious diseases and a comprehensive outbreak surveillance and response through an Integrated Disease Surveillance and Response program. [4]

6. Conclusions

From our point of view, Big Data is being indispensable for everything related to online content . At the moment, speaking as a whole, Big Data is used in large capacities, but it's usage will increase in the future. The content consists of information gathered so far, and research does not stop there, which leads us to think about Big Data.

Do not overlook the need to manage the performance of your data!

Big Data demonstrates that we are able to use more data than ever before in a faster rate of speed than was possible in the past. This ability to get multiple perspectives is a huge benefit. However, if the data is not managed in an efficient manner, it will cause serious problems for the enterprise. Therefore, we need to build flexibility on our path to build our Big Data plan.

References

[1] Brand, W., 2013. *Big Data For Dummies*, New Jersey: John Wiley & Sons;

[2] Frampton, M., 2014. *Big Data made easy*, New York: Apress;

[3] Berman, J., 2013. *Principles of BigData*, Waltham: Elsevier;

[4] McKinsey Global Institute, 2011.*Big data: The next frontier for innovation,competition, and productivity*;

[5] *Oracle Big Data Strategy Guide*, http://www.oracle.com/us/technologies/big-data/big-data-strategy-guide-1536569.pdf

Radu Ioan ENACHE graduated in 2013 from the *Economic Marketing Faculty* at the *Academy of Economic Studies* in Bucharest. At the moment he is pursuing the *Database for Business Support* master program. His area of interest are: Mobile Apps, Web Development and website optimization.

Marian Adrian ENE graduated in 2013 from the *Economic Marketing Faculty* at the *Academy of Economic Studies* in Bucharest. At the moment he is pursuing the *Database for Business Support* master program. His area of interest are: Databases, Web Development and multimedia applications.

Customer Data Analysis Model using Business Intelligence Tools in Telecommunication Companies

Monica LIA
University of Economic Studies, Bucharest, Romania
monica.lia@gmail.com

This article presents a customer data analysis model in a telecommunication company and business intelligence tools for data modelling, transforming, data visualization and dynamic reports building . For a mature market, knowing the information inside the data and making forecast for strategic decision become more important in Romanian Market. Business Intelligence tools are used in business organization as support for decision making.
Keywords: *Customer Analysis, Business Intelligence, Data Warehouse, Data Mining, decisions, self-service reports, interactive visual analysis, and dynamic dashboards, Use Cases Diagram, Process Modelling, Logical Data Model, Data Mart, ETL, Star Schema, OLAP, Data Universes*

1 Introduction

Business Intelligence tools refer to those software applications designed to retrieve, analyse or report data. In business intelligence tools are included a wide kind of applications: spreadsheets, visual analytics, querying software, data mining software, and data warehousing software or decision support software. A business intelligence platforms brings together a different kind of business intelligence tools which have the final scope support decision making at all levels in economic organisation.

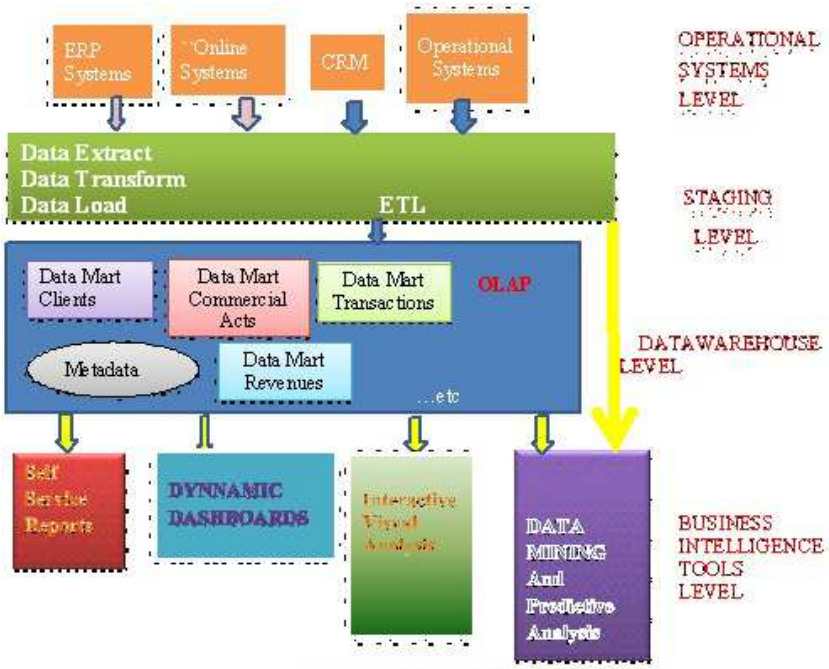

Fig. 1. Business Intelligence Platform

Modern Business Intelligence platform should provide an end-to-end infrastructure, solutions and technologies that support following issues: information integration, master data management, data warehousing, BI tools. A business intelligence platform includes the four levels described in the figure 1: operational systems level from which the data is collected, staging level for extracting,

loading and transformation of data for modelling in data warehouse. The last level is represented by business intelligence tools used for decision making.
Data Mining means predicting the future based on analysing information from the own systems.

Data Mining is made on large sets of data from different data sources and include four stages Exploration, Model building and validation; Deployment, Reports preparation.

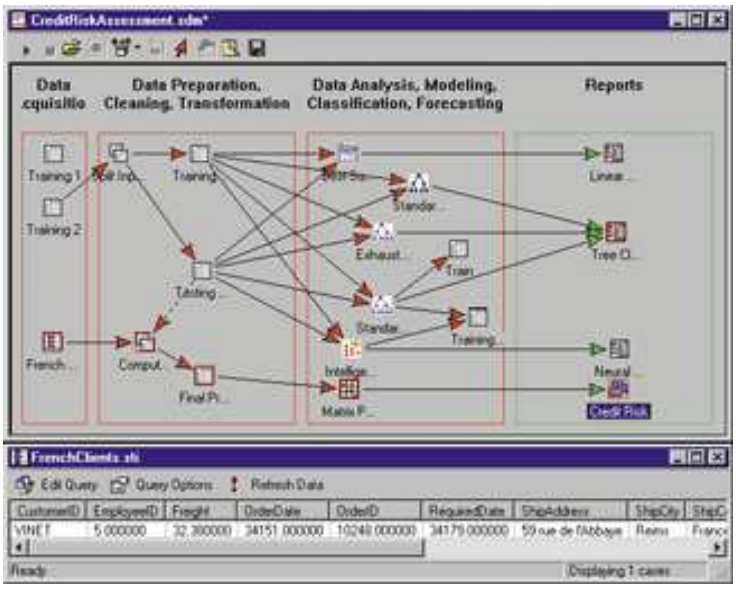

Fig. 2. Data Mining Stages. Source http://www.statsoft.com/

Data Mining could be made using data from staging level or from the data warehouse directly. Examples for Data Mining Business Intelligence tools are SAS Enterprise Miner, IBM SPSS, and Business Analytics for Information Builders. Data Mining could be made also without these tools. Looking for what is inside the data is a beautiful and difficult job.

Self Service Reports

Self-service business intelligence means that business users can create their own reports without IT department help. Usually this is possible it after the staging level. The data is organized using OLAP technology, on Data warehouse level. For example the Business Objects universes is a business representation of organization's data that helps end users access data autonomously using common business terms and it isolates business users from the technical details of the databases where source data is stored.

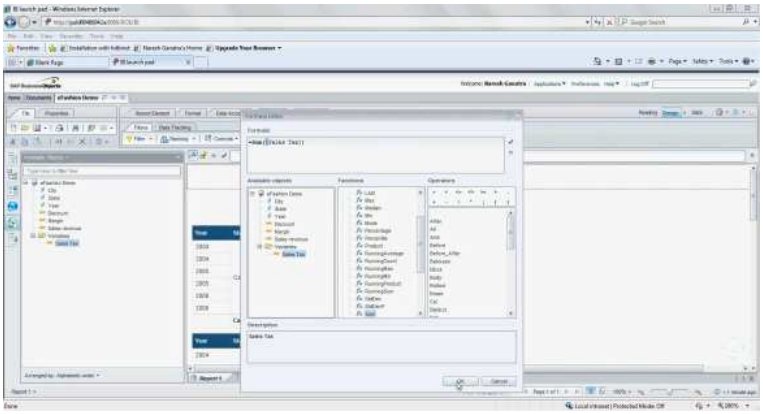

Fig. 3. Example of Self Service Report. Source http://datawarehouse4u.info/

Dynamic Dashboards

A way to organize together and manage multiple charts regarding on the same subject of interests is on dashboards. If the information in dashboards is not static and can be changed based on parameters values selections, those dashboards are called dynamic. In business is very often used because the information came from different sources and the volume of data is huge. Dynamic Dashboards are preferred by intermediate level of managers for giving a quick image by their business segment.

Fig. 4. Example of Dynamic Dashboard, Source, http://kb.tableau.com/articles/knowledgebase

Interactive Visual Analysis

Interactive Visual Analysis (IVA) is new part of business intelligence tool. The interactive visual analysis appears as a need for analysing high-dimensional data that has a large number of data points. Simple graphing reports without interactive techniques give an insufficient understanding of what is inside the data. complex datasets...

Using interactive visual analysis the user correlated views and iteratively select and examining features. The objective of analysis is to obtain knowledge which is not apparent from ordinary report. For Interactive Visual Analysis is important the perceptive and cognitive capabilities of humans who use it. This is necessary in order to extract knowledge from large and

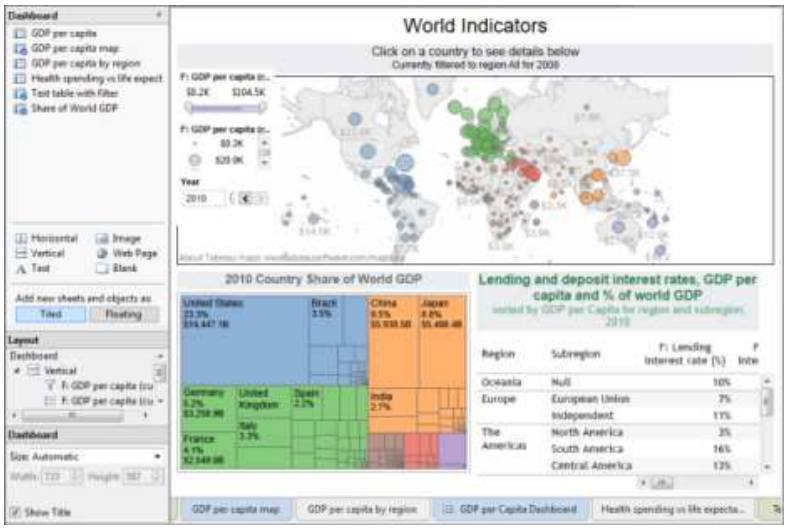

Fig. 5. Example of Interactive Visual Analysis, Source, *http://blog.activestrategy.com/performance-management-software-blog/*

Business intelligence tools are very useful for large companies. As example, the Telecom Market is very dynamic. Companies are looking for solutions for take advantage in this difficult competition. The analysis of what is inside the data from systems like traffic, sales, online, accounting become more important. The data from operational systems contains information about the client and how to keep this client, how to offer solutions for giving a better price or a better service. Also, gives ideas to decision makers on how to improve communication to client, how to improve network qualities and so one. In this paper the author presents a model of data analysis of a telecommunication customer.

2. Data Analysis Model

The customer data analysis model starts with analysis of which elements impacts customers' behavior. First, is clear understood that to become a company's client, the actor has to sign a contract. The contract is a result of an offer made by a company. To support the offers company has her own costs and stocks of products. The company gives to the client, on the offer base, services and/or products. Using company's services the client make voice calls, traffic on internet, content usage, transactions. All this traffic is made using the company's network. If the client needs assistance from the company has to interact using Interactions services (like IVR, Customer Care calls) or using Care Services for problems with devices. All this elements are presented using use case diagram in the figure 7.

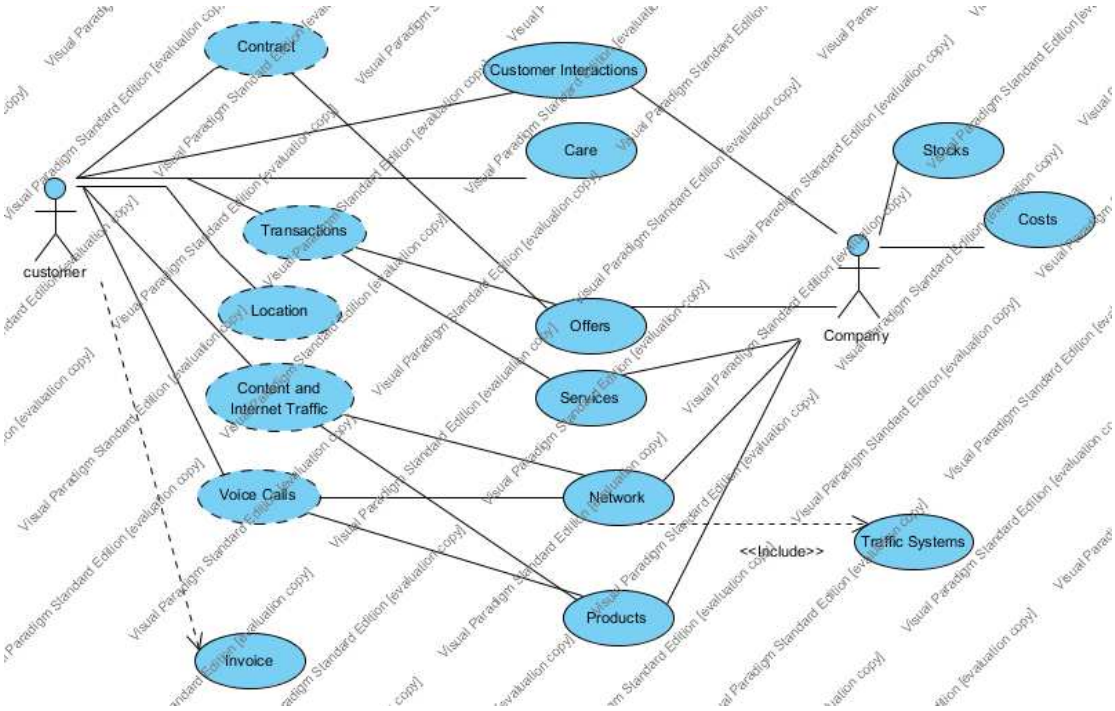

Fig.7. USE CASES DIAGRAM for Customer Analysis

The use cases diagram helps the understanding of what are the elements which determine customer's behaviour. This is important for understanding which data needs to be modelled in order to developed dynamic reports necessary for customer analysis. In a telecom company, information from customers comes from different data sources as: operational systems for customer's traffic, operational systems for contracts, invoices, and online systems. The data from operational systems are load in Data Warehouses. An example is presented in the Figure 8.

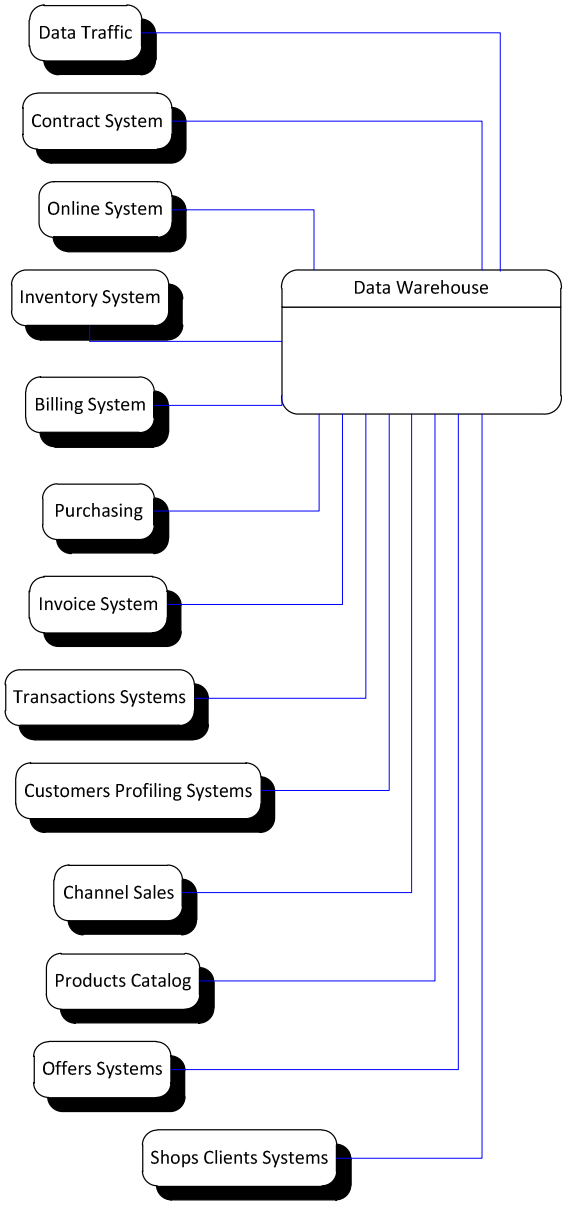

Fig.8. Data Warehouse Structure in Telecommunication Companies

All the information from operational systems needs to be modelled in order to make possible the customer data analysis.
After business process understanding, the logical business model must to be developed. Logical model is necessary for understanding how data will be modelled. For customer analysis the logical model is discuss with business owner and the sponsors of customer analysis project. The analysis is made by business analyst. For customer analysis the logical model is presented in the figure 9.

Based on logical Model is obvious now the large kind of analysis which can be made on customer. In the Table are some examples.

Table 1. Some Possible Customer Analysis

	Kind of Analysis Type
1	Customer Analysis per traffic and customer type
2	Customer with Smartphones Data Traffic Analysis
3	Smartphone Sales per channel distribution
4	Traffic analysis per acquisition channel
5	Geographical repartition of customer per volume of data used and product type
6	Offers and contract type evolution in period per channel

Customer Analysis per Traffic and Customer Type is possible if the Data Warehouse contains information about traffic and customers. Next we will define the steps to implement traffic events in Data Warehouse. This will be made in four steps: ETL stage, Data Marts, Universes Building, and Data Presentation in Dynamic Reports.

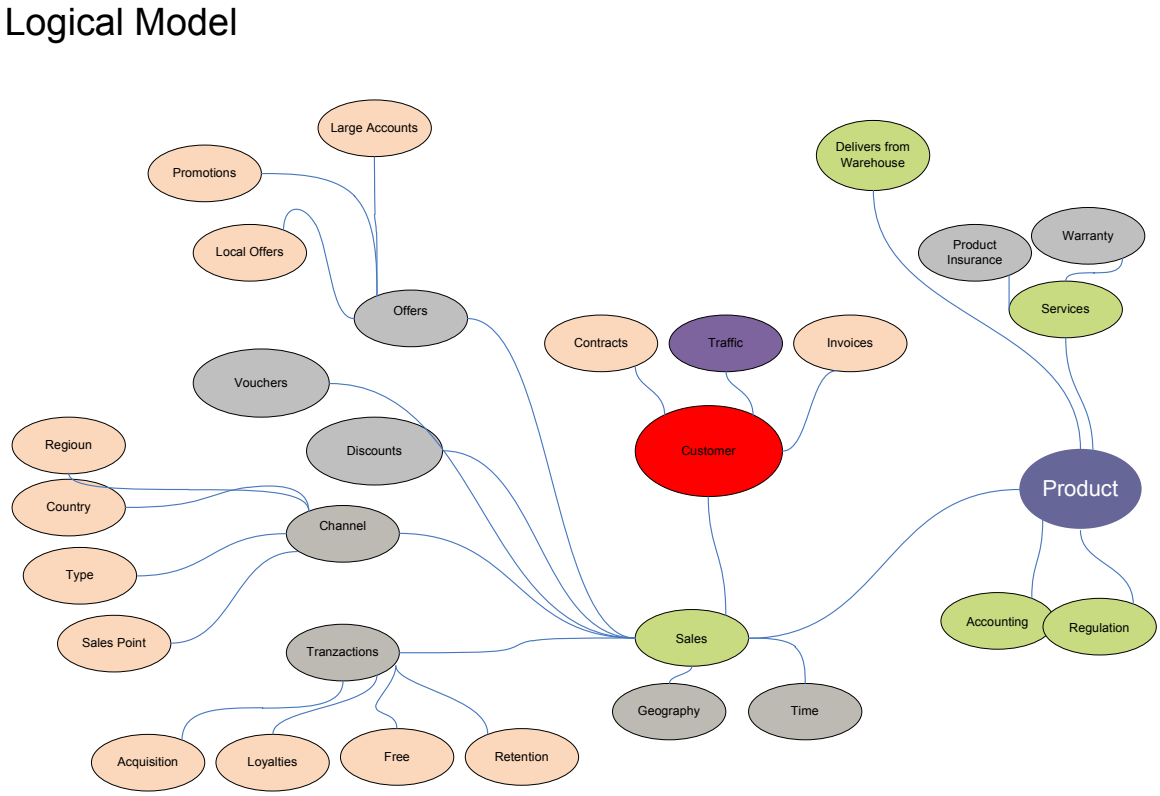

Fig.9. Logical Model for Customer Analysis

In ETL Stage Data of Operational Systems are Extract, Load and Transformed for Data Warehouse. Operational systems involves in traffic events are described in figure 10.

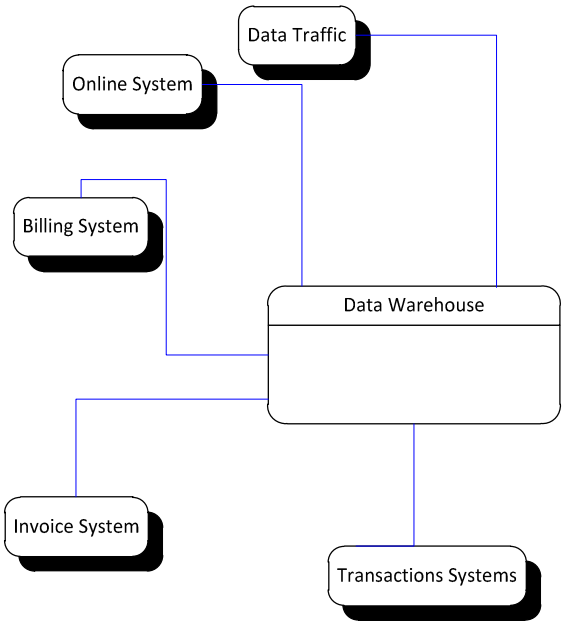

Fig. 10. Traffic Star Schema

The traffic CDR's must be load in Data Warehouse. The ETL is made with a special tool. One tool that can be used is Informatica. Data loading workflow for traffic CDR's from operational systems presents is described in figure 11.

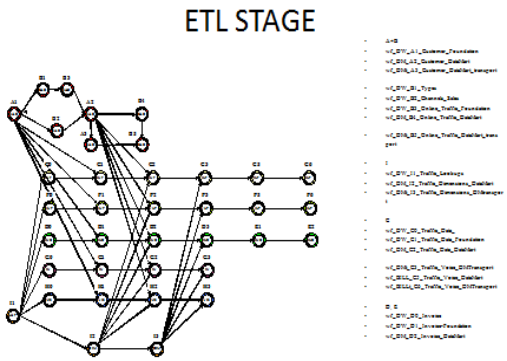

Fig. 11. Traffic ETL Stage Workflow

After data loading in stage, the data will modeled for data mart. Based on Data Mart can be construct cubes for OLAP analysis (base for data mining). The Data Mart model based on information in traffic CDR's is presented in the figure 12.

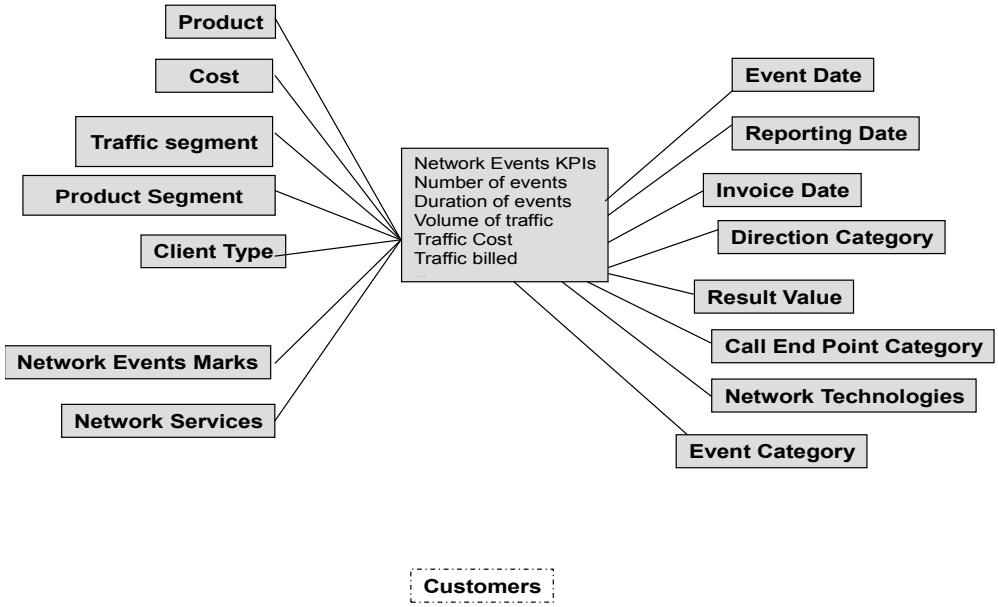

Fig.12. Traffic Star Schema Data Mart

The business needs for reporting are reflected in the star schema and also in data dimensions tables. In figure 13 are presented a Data Mart Traffic Dimension Tables.

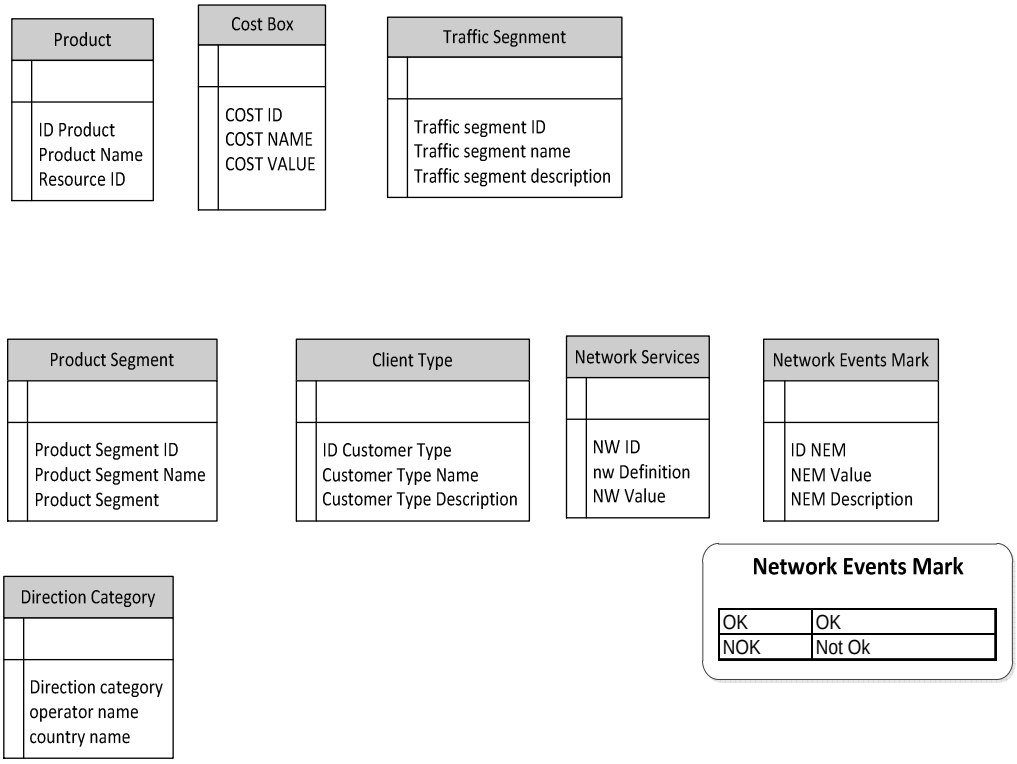

Fig.13. Traffic Data Mart Dimension Tables

The data modeled in Data Mart can be used with an OLAP business intelligence tool for Dynamic Reports Building. One example is SAP Business Objects InfoView. The Information is presented in dimensions and measures. We have one example presented in the figure 14.

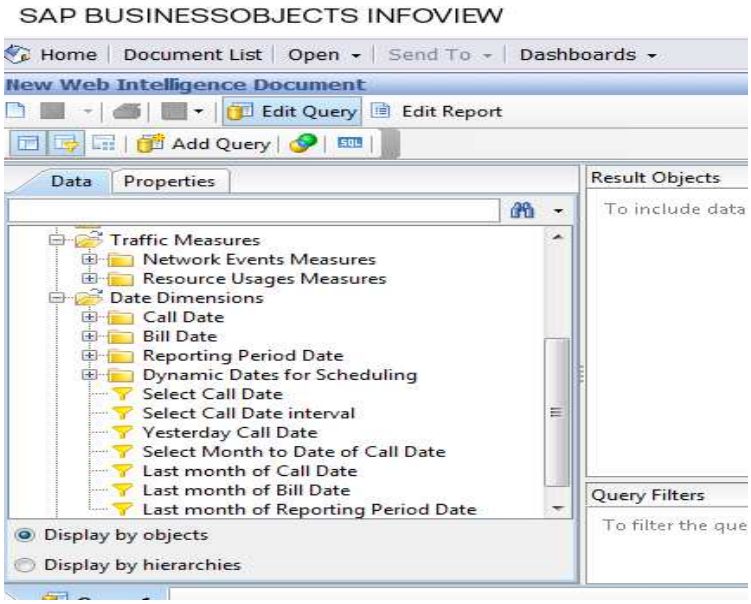

Fig.14. Traffic Star Schema

Using the data modeled in universes, users can develop self services reports. They can select information from traffic dimensions and correlate with customer's information. This is made to analyze customer's behavior by traffic components.

All the elements presented in logical model needs to be modeled in Data Warehouses and presented in universes. The modelling process will be made following the steps presented in this paper. In this way the final users will have all the elements for a complete analysis of customer's behavior.

3. Conclusions

The analysis of what is inside the data are base for sales forecast. Also, the future offers, services and products are adjust based on these dynamic reports. The importance of analyzing the own data about personal customers is very important to telecom companies and is easy to made using business intelligence tools. In this paper I presented a brief overview how to model the data from operational systems in order to help the final users to develop self-service and dynamic reports.

Acknowledgments

Personal thanks to professor Ion Lungu who help me in my odyssey of performing. And also to my kids, all special, who let me write in the nights. They sleep so well so I can be very concentrate on my work...

References

[1] Sid Adelman, Larissa Terpeluk Moss, *Data Warehouse Project Management*, Ed. Addison-Wesley, Boston, 2004.

[2] John Wang, *Data Warehousing and Mining, Concepts, Methodologies, Tools and Applications*, Ed. Information Science Reference, New York, 2004.

[3] Ralph Kimball, Margy Ross, *The Data Warehouse Toolkit, The Complete Guide to Dimensional Modelling, Second Edition*

[4] I. Lungu, A. Bara, *Sisteme informaticeexecutive*, ASE Publishing House, 2007

[5] http://www.statsoft.com/

[6] http://datawarehouse4u.info/

[7] http://kb.tableau.com/articles/knowledgebase

[8] http://blog.activestrategy.com/performance-management-software-blog/

Monica MANEA graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 1996. She has a postgraduate diploma in Accounting and Management Information Systems. She attended Stanford University Project Management Classes between 2005-2006. At present she is studying for the doctor's degree at the Academy of Economic Studies Bucharest.

Business Intelligence Methods for Sustainable Development of the Railways

Aida-Maria POPA
Academy of Economic Studies, Bucharest, Romania
aida_popa@yahoo.com

This paper aims to present a new approach of business intelligence technologies in the context of sustainable development of the railways. The concept of business intelligence is increasingly used in the developed companies and considering that the current economic market is more dynamic from year to year, business intelligence solutions plays an important role for companies to be able to develop efficient plans for both short-term and medium and long term developing. This paper will focus on two technologies: data-warehouse and data-mining and how are they use in the railway business. The subject adapts to the current development trend of European countries to direct the transport of freight and passengers to the railway for support environment.
Keywords: *sustainable development, business intelligence, data-warehouse, data-mining*

1 Introduction

Railway has demonstrated in the last two centuries an important contribution to the prosperity of states, to the development of national and global economies but also to the possibility of transportation in any point on the planet.
By its nature, the railway is composed of complex processes and activities which are conducted continuously while on a network having significant geographical spread. It involves humans, materials and financial resources with unpredictable events, hence the need for strong systems to respond with flexibility and accuracy to determine market requirements and increase the quality of services offered to all customer`s categories.
The evidence of tradition of Informatics Romanian Railway dates from July 1, 1967 when it was founded the Informatics Centre for Rail Transport (then called Electronic mechanized Computing Center Ministry of Railways). In short time, it has succeeded in establishing itself as a strong core and it managed to maintain and strengthen this status even in the present form of reorganization and transition.
Through activity carried out over the years since it was founded, the Informatics Centre for Rail Transport consistently succeeded to bring beneficial solutions to its customers by deploying applications and advanced information systems. An important aspect of these systems is to lower costs due the fact that they were made in the railway sector, which brought another benefit that is a closely knowledge about processes and activities that were to be computerized.
Currently, the main objective of Romanian Railways is to develop business activity by increasing efficiency and ensuring a future of rail transport, keeping up with modern times and considering the ever present main competitor in the transport field, namely road transport. To achieve this objective, the Romanian Railways is going through a restructuring process that is based on substantial investments.

2. Some aspects regarding Business Intelligence

According to the Romanian Association of Economic Intelligence, business intelligence represents all the activities of research, collecting, processing and information dissemination of useful economic agents in order to gain competitive advantage by exploiting them in the defensive and / or offensive manner.
The paper "Business Intelligence Roadmap" [1] presents business intelligence as an architecture and a collection of applications and operational integrated databases and decision support systems, which provide the

business community easier access to business data.

According to Brândaș C. [2] the most important objectives of Business Intelligence are:

- Gathering and analyzing large volumes of data and information extracted from both the operational database and the data warehouse within the organization;
- The combination of two processes, the knowledge management and the decision management;
- Obtaining complex information for managers and competitive advantage by exploiting technologies to support decision making within the organization.

Website [3] defines Business Intelligence (BI) as a process that aims to serve and support business process management. Streamlining the process through precise timely and correctly founded decisions is the main purpose of the choice of implementing a Business Intelligence system. This process consists people, a set of tools and specialized software applications, methods, services and techniques, all of which been necessary for collecting, storage, analysis and better use of data and information derived from the other processes that occur within the organization. Operation process and also business decision making, depends entirely on the aforementioned elements.

Architectural principles of a business intelligence system are [4]:

• scalability and high performance;
• complete functionality;
• continuous development
• openness and extensibility;
• developing and fast running.

2. Sustainable development

2.1. Definition of the concept of sustainable development

What sustainability means is still no universally agreed definition on. There are a variety of views on what it is and how it can be obtained. The concept of sustainability stems from the idea of sustainable development which has become a usual terminology at the World's first Earth Summit in Rio in 1992 (Brundtland Report for the World Commission on Environment and Development).

The original definition of sustainable development is usually considered to be:

"Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs. It contains within it two key concepts:

- the concept of needs, in particular the essential needs of the world's poor, to which overriding priority should be given; and
- the idea of limitations imposed by the state of technology and social organization on the environment's ability to meet present and future needs."[5][6]

The concept of sustainable development can be characterized in many different ways, through the following elements:

- Economics: efficiency, growth, stability;
- Society: standard of living, equity, social dialogue and delegation of responsibilities, the protection of culture / heritage;
- Ecology: conservation and protection of natural resources, biodiversity, avoiding pollution.

Development can be considered sustainable when it meets together economic, social and environment objectives.

Sustainable development is a fundamental objective for the European Union. The purpose for which it addressed this concept is continually improving the quality of life of present and future prosperity, an approach that takes into account the integration of economic development, environment and society.

The purpose for which it addressed this concept is continually the quality of life, improving present and future prosperity, an approach that takes into account the

integration of economic development, environment and society.

2.2. The advantages of rail transport

Sustainable development shows its concern for saving energy, air pollution and numerous traffic accidents. Therefore trends were followed for reinstatement on the main plan of rail transportation, in this way being advancing the development of high-speed rail between cities and the corridors with a high population density.

Many international studies have shown clearly that railway is the best way to transport relating to protect the environment.

In Figure 1 highlights the advantages of rail versus other transport systems against four important aspects: energy consumption, exhaust emissions, traffic safety and the land surface required infrastructure.

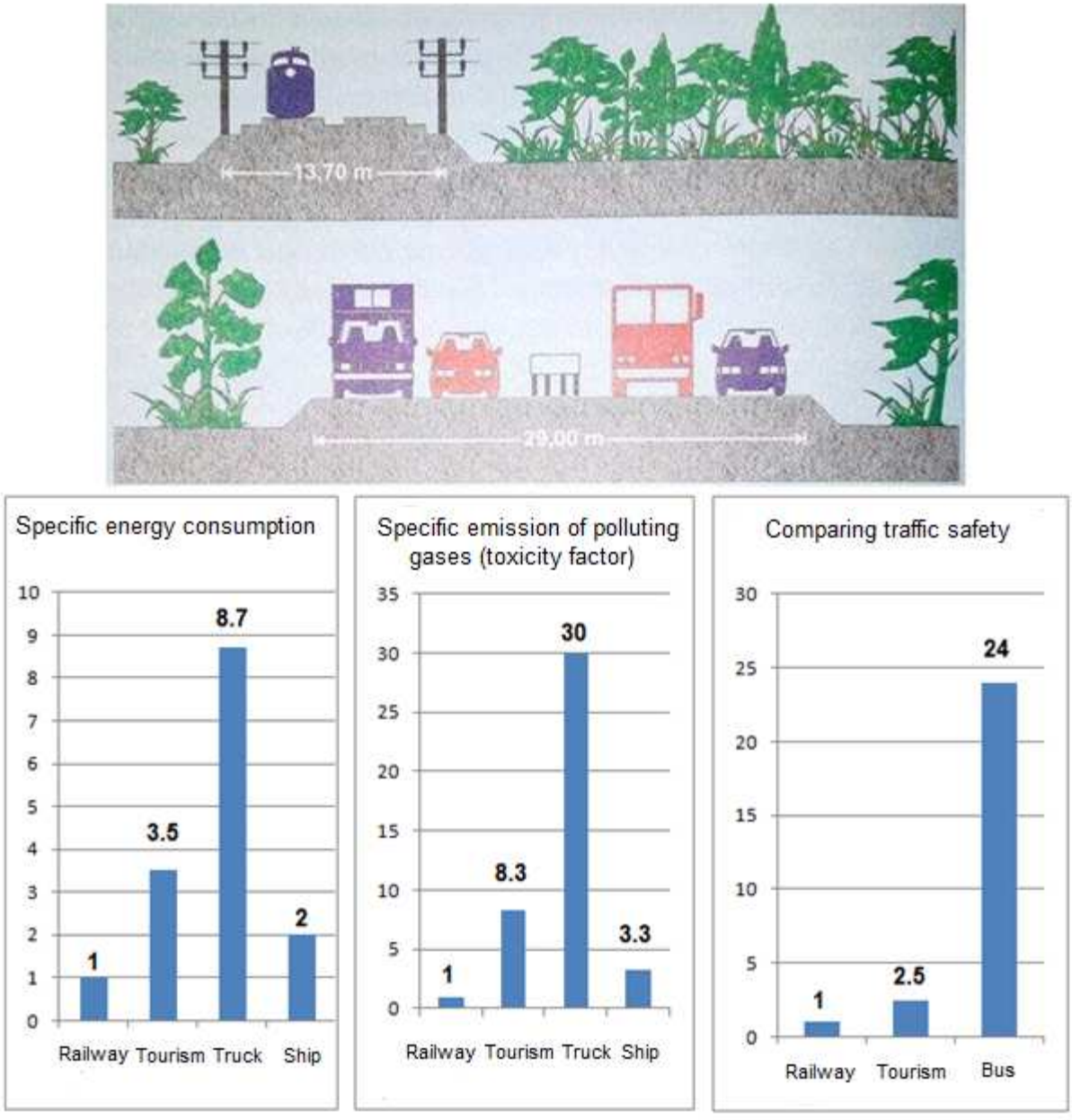

Fig. 1. The advantages of rail transport compared with other transport systems [7]

In terms of environmental pollution we clearly observed advantages of railway transport compared with other systems, the concentration of pollutants is even up to 30 times lower than the concentration of a truck emissions, in the case of freight transport.

By comparing energy consumption registered at various ways of transport, we see a consumption up to 3.5 times higher in case of transport by road and twice as high in case of ship transport compared to rail passenger transport. For freight transport, consumption energy registered by trucks becomes 8.7 times higher than consumption energy registered by rails.

The highest degree of traffic safety is registered by rail and air transport. If we refer to a similar passenger transport capacity, accident risk of the bus is 24 times higher than rail transport. However, in Romania, the risks of road transport is far outweigh of value from the graphic.

Given the fact that Romania belongs in the category of medium populated states, with a population density of about 97 inhabitants per km², land areas are difficult to access for developing transport infrastructure. In case of a modern railway are used about 14 meters wide surface transport compared to the same highway transport capacity that will have 4 lanes and will occupy about 31.5 meters width. Another disadvantage of road transport is also the material that is used to make highway which is not always good for the environment.

3. Data-Mining

According to [8] is presented a methodology called "Methodology for Railway Demand Forecasting Using Data Mining ". Considering the genuine complexity of processes of Knowledge Discovery in Databases (KDD), it was developed this methodology which is based on principles of activity planning. The steps belongs to the knowledge discovery process are set before being executed taking into account the objectives of each KDD request. The application of the methodology is split in four phases as is illustrated by Figure 2.

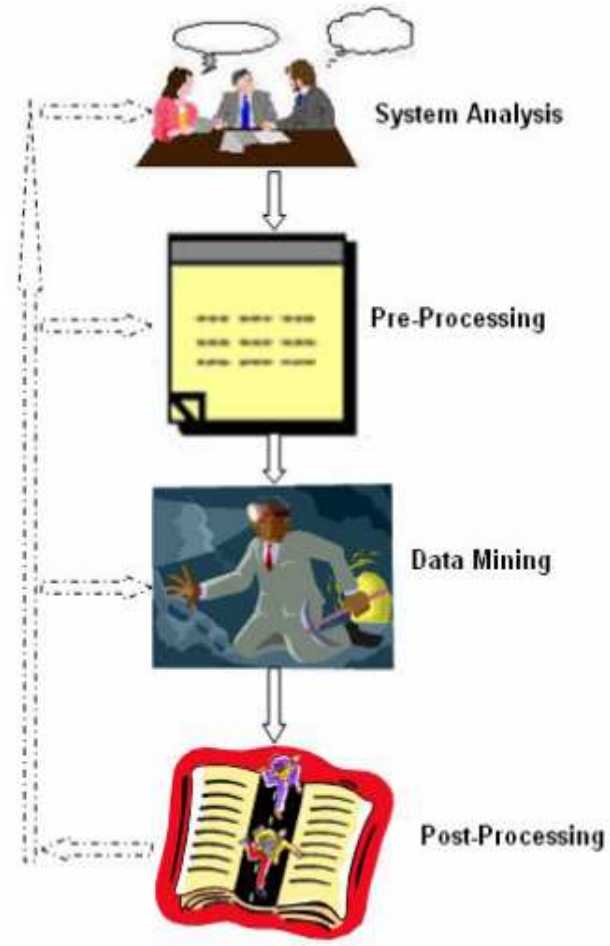

Fig.2. Methodology for Railway Demand Forecasting Using Data Mining [8]

This methodology conjures that is an interactive and iterative process. This translates into the fact that the KDD analysts can return to every step previously browsed if they want to find better results, other than those was already discovered. To achieve this requirement, it is required a precise and detailed documentation about developed actions and achieved results. In this sense it is preferred to use documentation models to have the opportunity to choose which procedures to be adopted taking into account the number and diversity of situations and possibilities.

The phases of methodology are:

a) System Analysis (Figure 3) is the first phase of methodology. The most important

objective of this phase is to define the types of requests that have to be performed by applying techniques of KDD process. The wish in this regard is to identify process objectives and also their resolving or improving. The activities contained in the system analysis phase are: definition of the actors, description of the problem, definition of the objective, expectations and deadline.

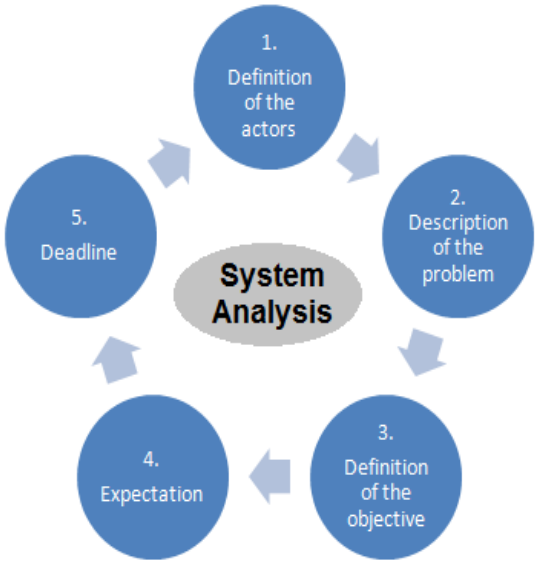

Fig.3. System analysis activities

b) Pre-Processing (Figure 4) is the second phase of methodology. Capturing, organizing, treating and preparing are the functions that make up this phase. These functions are predecessors to Data Mining phase, having a colossal importance for the knowledge discovery process.

After defining the desired outcomes it is going the first activity, choice of the technique, after which we have to choose which techniques can be used that are more close to obtaining results with higher precision.

The activity of selection of the data is indispensable for the pre-processing phase. The selection is necessary for reporting the origin of the information regardless of their source (transactional or from data-warehouse).

Cleaning the data is an optional activity, being used only in case of absent information, inconsistencies, and values that are not pertaining to the domain. According to Kimball and Ross [9] in case of using the information from a data-warehouse the possibility and necessity of cleaning is lower because the creation of a data-warehouse has a process when the database is cleaning.

Codification is the pre-processing activity on which depends the data representation during the KDD process.

The activity of normalization is achieved by assigning a new range to an attribute so that the values could be within the new range in a specified interval (for example, from -1.0 to 1.0 or from 0.0 to 1.0). The adjustments become needful especially for preventing some of the attributes from having a larger range of values than others, for not influence the tendencies of the algorithms of Data Mining that are used.

Enrichment represents the capacity to add additional information to the already recorded data so that these provide more entities to the knowledge discovery process.

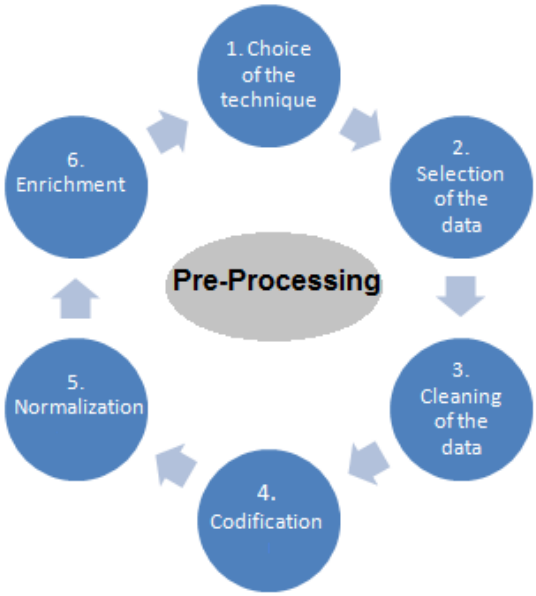

Fig.4. Pre-processing activities

c) Data Mining (Figure 5) is the principal phase of the proposed methodology having the role in the helpful search for new and useful knowledge obtained from the data. This is why the Data Mining and the KDD process are referred by many authors as

two terms that are synonymous. The activities contained in the Data Mining phase are: partition of the data, choice of the tool and data mining.

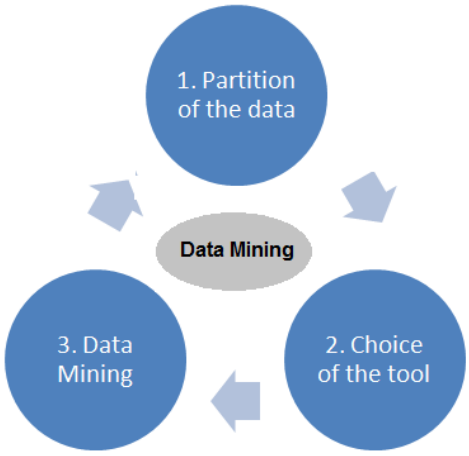

Fig.4. Data-mining activities

d) Post-Processing (Figure 6) is the last phase of the methodology. It includes the simplification and presentation activities of knowledge models achieved in the Data Mining phase. In this phase the results obtained are evaluated and are defined new alternatives of data surveys [8].

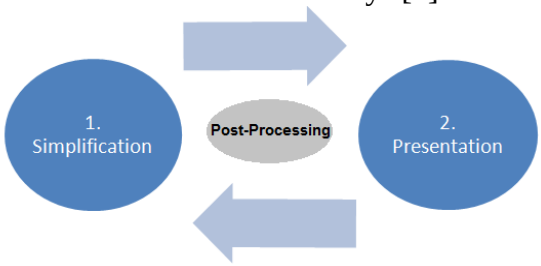

Fig.5. Post-processing activities

4. Data-Warehouse

The data warehouse represents a set with a huge volume of data (reaching up to terabytes) that is used as a data source that is compact, integrated and complete and it is a base for a variety of types of information systems (for example: decision support, executive, business intelligence) aiming to provide stored data from noteworthy sources (operational data, external files, archives, etc.) used at supporting the decision-making process in a company business [10].

Data warehouse represents a technology with an increasingly present in business being increasingly appreciated in the last years, becoming a performing solution in terms of clients and business development of a company. The fact that this technology is becoming more prevalent in more and more activity fields reported an efficiency of a large variety of operations and an improvement of market intelligence [11].

Data warehouses comprise data coming from many different information sources that are converted for a multidimensional representation used by Business Intelligence Systems [12].

Data warehouse represents a more complex form of database with a large volume of data, usually designed using the traditional relational model, which contains many historical data of a certain interest [4].

Data warehouse is organized as a unique source of data and information for the entire organization which represents a fundamental principle of the integrity, where data is stored in a single, common form of representation of data from all sources (databases, external files, archives, etc.) settling specifically conventions on the designation fields, coding systems, representation of measure units, representation mode for calendar data, avoiding duplication of the same thing originated by different sources (departments) [13].

Figure 6 illustrates a model for data-warehouse tables organisation for railway level crossing. Is observed the higher number of atributes which are the base for a detailed data storage.

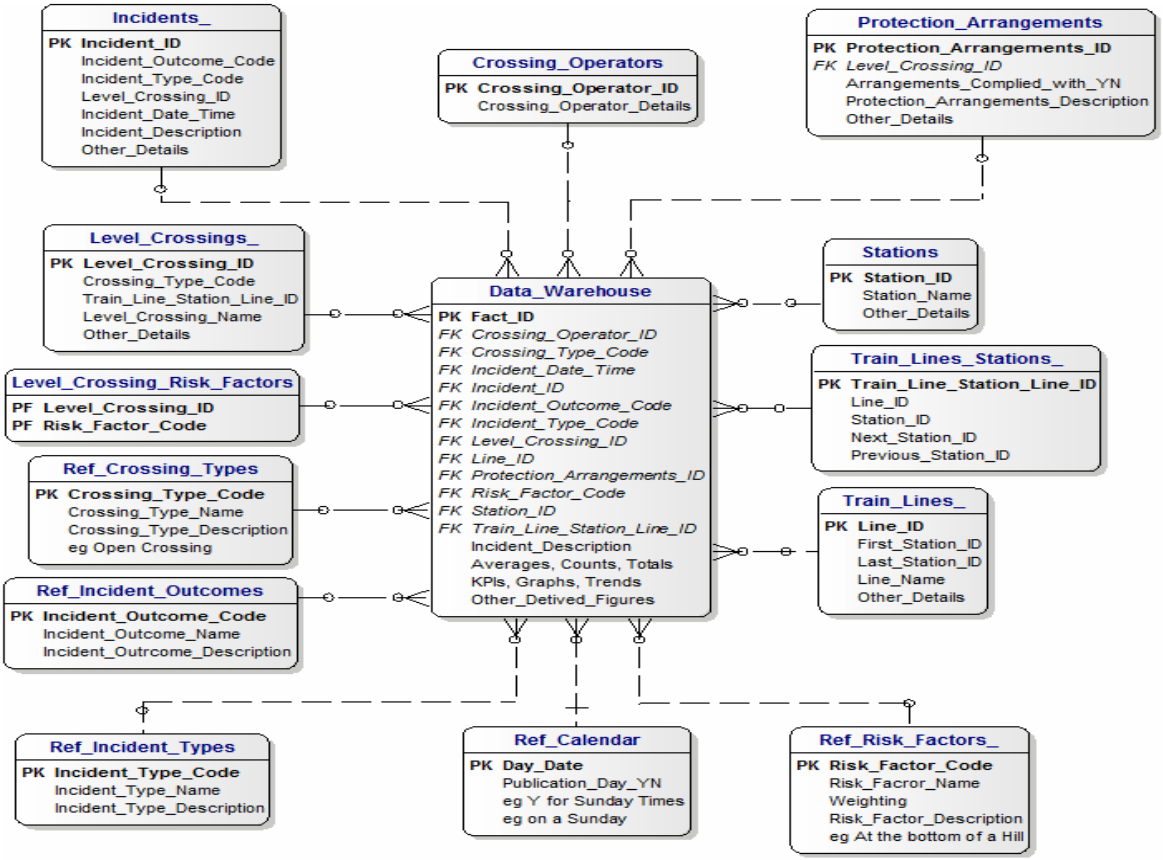

Fig. 6. Data Warehouse for Railway Level Crossings [14]

Conclusions

More and more European countries are currently favors for the transport of freight and passengers on the railroad for support environment. In this context, Romania is trying to adapt to this trend by developing regional operational programs. Business intelligence systems are the basis for adapting to this trend because it showed his usefulness in many developed or developing countries. The shift to the modal or mixed transport requires a large volume of records to observe how it can be constantly improved. It is also important to know how the data are used for the discovery of knowledge and supporting decision-making at the CFR. In conclusion we can notice that IT technologies (data-mining and data-warehouse) are successfully adapts to this area.

For the applicative part, there was presented a star schema of data-warehouse for railway level crossing where the data is stored in many tables having a very high level of detail. These data are used by business intelligence system to achievement reports, forecasts and support the management in making beneficial decisions for the company.

References

[1] Larissa T. Moss, Shaku Atre - *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*, 2003, ISBN 978-0201784206.

[2] Claudiu Brândaș - *Contribuții la conceperea, proiectarea și realizarea sistemelor suport de decizie*, Teză de doctorat Universitatea „Babeș-Bolyai”, Cluj-Napoca, 2007.

[3] http://www.comunitateaerp.ro/utile/28

[4] B. Nedelcu, “*Business Intelligence Systems*”, Database Systems Journal, Vol. IV, Issue 4/2013, pg. 12-20, 2013, ISSN 2069-3230. Available:

http://www.dbjournal.ro/archive/14/14_2.pdf
[5] http://www.globalfootprints.org/sustainability
[6] International Institute for Sustainable Development https://www.iisd.org/sd/
[7] Viorel Simuț - *Managementul transportului feroviar*, Editura Asab, București , 2001, ISBN 973-85247-0-9.
[8] http://www2.sas.com/proceedings/forum2007/161-2007.pdf
[9] Kimball, Ralph, Ross, Margy. *The Data Warehouse Toolkit. Guia Completo para Modelagem Dimensional*, Editora Campus. Rio de Janeiro, 2002.
[10] Popa Aida Maria – *Data Warehouse Pyramidal Schema Architecture – Support for Business Intelligence Systems*, The 14th International Conference on Informatics in Economy, 30 aprilie – 3 mai 2015, Proceedings if The 14Th International Conference of Informatics in Economy, Bucuresti, ISSN 2284-7472.
[11] John Foley, **"***The Top 10 Trends in Data Warehousing*", March 10, 2014. Available: http://www.forbes.com/sites/oracle/2014/03/10/the-top-10-trends-in-data-warehousing/
[12] G. Satyanarayana Reddy, M. Poorna Chander Rao, R. Srinivasu, S. Reddy Rikkula, "*Data Warehousing, Data Mining, OLAP and OLTP Technologies Are Essential Elements To Support Decision-Making Process in Industries*", International Journal on Computer Science and Engineering(IJCSE), vol. 2, No. 9, pp. 2865-2873, 2010, ISSN 2865-2873.
[13] M. Velicanu, Gh. Matei, *Tehnologia inteligenta afacerii*, Editura Ase, Colectia Informatica, Bucuresti, 2010, ISBN 978-606-505-311-3.
[14]http://www.databaseanswers.org/data_models/railway_level_crossings/data_warehouse.htm

Aida Maria POPA graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2012. She graduated from the Economic Informatics Master of the Academy of Economic Studies in 2014. Currently, she is a PhD candidate, coordinated by Professor Manole VALICANU in the field of Economic Informatics at the Bucharest University of Economic Studies. Her scientific fields of interest include: Databases, Data Warehouses, Business Intelligence, Decision Support Systems and Data Mining.

Stochastic Processes and Queueing Theory used in Cloud Computer Performance Simulations

Florin-Cătălin ENACHE
Bucharest University of Economic Studies
catalin.enache@live.com

The growing character of the cloud business has manifested exponentially in the last 5 years. The capacity managers need to concentrate on a practical way to simulate the random demands a cloud infrastructure could face, even if there are not too many mathematical tools to simulate such demands.This paper presents an introduction into the most important stochastic processes and queueing theory concepts used for modeling computer performance. Moreover, it shows the cases where such concepts are applicable and when not, using clear programming examples on how to simulate a queue, and how to use and validate a simulation, when there are no mathematical concepts to back it up.
Keywords*: capacity planning, capacity management, queueing theory, statistics, metrics*

1 Introduction During the last years, the types and complexity of people's needs increased fast. In order to face all changes, the technology had to develop new ways to fulfill the new demands. Therefore, I take a deeper look into the basic terms needed for understanding the stochastic analysis and the queueing theory approaches for computers performance models. The most important distribution for analyzing computer performance models is the exponential distribution, while the most representative distribution for statistical analysis is the Gaussian (or normal) distribution. For the purpose of this article, an overview of the exponential distribution will be discussed.

2.1 The Poisson Process

In probability theory, a Poisson process is a stochastic process that counts the number of events and the time points at which these events occur in a given time interval. The time between each pair of consecutive events has an exponential distribution with parameter λ and each of these inter-arrival times is assumed independent of other inter-arrival times. Considering a process for which requests arrive at random, it turns out that the density function that describes that random process is exponential. This derivation will turn out to be extremely important for simulations, in particular for applications modeling computer performance. A typical example is modeling the arrival of requests at a server. The requests are coming from a large unknown population, but the rate of arrival,λ can be estimated as the number of arrivals in a given period of time. Since it is not reasonable to model the behavior of the individuals in the population sending the requests, it can be safely assumed that the requests are generated independently and at random.

Modeling such a process can help answering the question of how a system should be designed, in which requests arrive at random time points. If the system is busy, then the requests queue up, therefore, if the queue gets too long, the users might experience bad delays or request drops, if the buffers are not big enough. From a capacity planner point of view, it is important to know how build up a system that can handle requests that arrive at random and are unpredictable, except in a probability sense.

To understand and to simulate such a process, a better understanding of its randomness is required. For example, considering the following time axis (as in the second figure), the random arrivals can be represented as in the figure below.

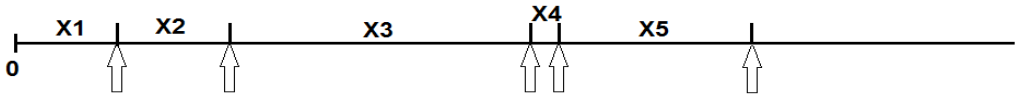

Fig.1. Random arrivals in time

If X is the random variable representing the times between two consecutive arrivals (arrows), according to the PASTA Theorem(Poisson Arrivals See Time Averages)[1], it is safe to assume that all X-es are probabilistically identical. Describing this randomness is equivalent to finding the density function of X that represents the time distance between two consecutive arrows.

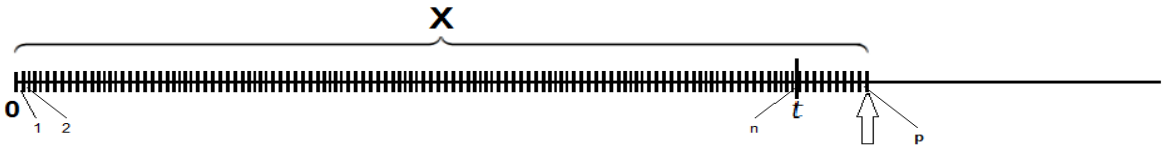

Fig.2. Interval of length t divided into n intervals.

The problem described above needs to be transformed so that it can be handled with known mathematical tools. Supposing that an arbitrary interval of length t is chosen, then the probability of the time until the first arrival is longer than t is P(X> t). This is by definition 1-$F_X(t)$, where $F_X(t)$ is the distribution function to be calculated. If time would be discrete, by dividing the interval between 0 and t into n intervals, the calculating $F_X(t)$ reduces to calculating the probability of no arrow in the first n intervals, and switching back to the continuous case by taking $n \to \infty$.

Let p be the probability that an arrow lands in any of the n time intervals, which is true for any of the n intervals since any of them is as likely as any other to have an arrow in it, then $P(X > t) = (1-p)^n$, which is the probability on no arrow, 1-p, in the first n intervals. As mentioned, when taking $n \to \infty$, $p \to 0^+$ and $np = \lambda t$. The equality $np = \lambda t$ represents the average number of arrows in n intervals – np – which is equal to the average number of arrows calculated as λt - the arrival rate multiplied by the length of the interval.After switching to the continuous case, it is derived that:

$$P(X>t) = \lim_{\substack{n\to\infty \\ p\to 0^+ \\ np=\lambda t}} (1-p)^n = \lim_{\substack{n\to\infty \\ t>0}} (1-\frac{\lambda t}{n})^n \underbrace{\qquad}_{\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x} e^{-\lambda t} \quad (1)$$

Which is equivalent to

$$P(X \le t) = 1 - P(X > t) = \begin{cases} 0, (t < 0) \\ 1 - e^{-\lambda t}, (t \ge 0) \end{cases}$$

$$\text{and } f_X(t) = \frac{d}{dt} F_X(t) = \begin{cases} 0, (t < 0) \\ \lambda e^{-\lambda t}, (t \ge 0) \end{cases} \quad (2)$$

2.2 The exponential distribution.

The random variable X derived from the Poisson process studied in section 2.1 of this paper is called exponential with the parameter λ (X~Exp(λ)).The probability density function(PDF) of X is defined as $f_X(t) = \begin{cases} 0, if t < 0 \\ e^{-\lambda t}, if t \ge 0 \end{cases}$, which plots as in the figure below for different values of the parameter λ.

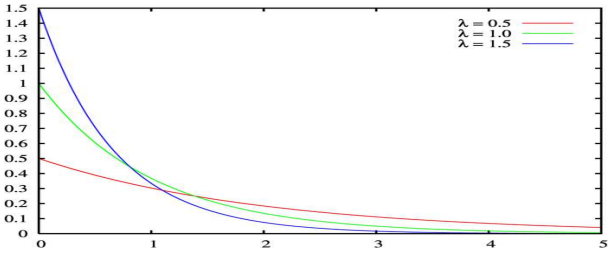

Fig.3. PDF for λ in (0.5, 1.0, 1.5)

Integrating by parts,it is easy to demonstrate the property that $\int_0^{\infty} \lambda e^{-\lambda t} dt = 1$, which is actually obvious, since the sum of all probabilities of a random variable X has to add up to 1.If X~Exp(λ) then the following properties are true[2] :

- The expected value the random variable X, $E(X)=\int_0^{\infty} t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}$ (3) ,
- Expected value of X^2 , $E(X^2)=\int_0^{\infty} t^2\lambda e^{-\lambda t} dt = \frac{2}{\lambda^2}$ (4) and
- The variance of X, $V(X)=E(X^2) - [E(X)]^2=\frac{2}{\lambda^2}-\left(\frac{1}{\lambda}\right)^2=\frac{1}{\lambda^2}$ (5) .

When used in simulating computer performance models,the parameter λ denotes usually the arrival rate. From the properties of the exponential distribution, it can be deduced that the higher the arrival rate λ is, the smaller are the expected value – E(X) – and variance – V(X) – of the exponentially distributed random variable X.

3.1. Introduction to the Queueing Theory M/G/1 Problem – FIFO Assumption

Considering a system where demands are coming are random, but the resources are limited,the classic queueing problem is how to describe the system as a function of random demands. Moreover, the service times of each request are also random, as in figure 4:

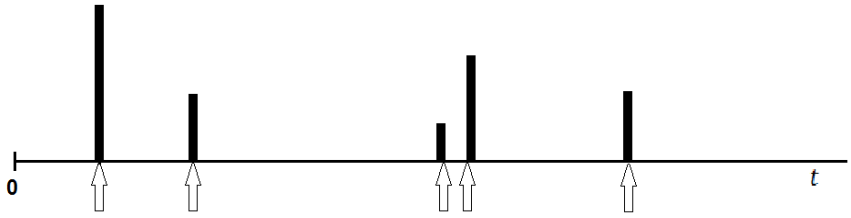

Fig. 4. Random arrivals with random service times

From a request point of view, when a new request arrives, it has two possibilities:

- It arrives and the server is available. Then it keeps the server busyfor a random amount of time until the request is processed, or
- Typical case, when a request arrives, it finds a queue in front of it, and it needs to wait.

The queueing theory helps answering questions like what is the average time that a request spends waiting in queue before it is serviced. The time a request must wait is equal to the sum of the service times for every request that is in the queue in front of the current request plus the remaining partial service time of the customer that

was in service at the time of the arrival of the current request.
Calculating the expected waiting time of the new requestmathematically, it would be the sum(further named "convolution") of the density functions of each of the service time requirements of the requests in the queue, which could be any number of convolutions, plus the convolution with the remaining partial service time of the customer that was in service at the time of the arrival of the current request. Furthermore, the number of terms in the convolution, meaning the number of requests waiting in the queue, is itself a random variable[1].

On the other side, looking at the time interval between the arrival and the leave of the n^{th}request, it helps in developing a recursive way of estimating the waiting times. The n^{th}request arrives at time $\mathbf{T_n}$ and, in general, it waits for a certain amount of time – noted in the below figure with $\mathbf{W_n}$. This will be 0 if the request arrives when the server is idle, because the request is being served immediately. To enforce the need of queueing theory, in real-life, a request arrives typically when the server is busy, and it has to wait. After waiting, the request gets serviced for a length of time $\mathbf{X_n}$, and then leaves the system.

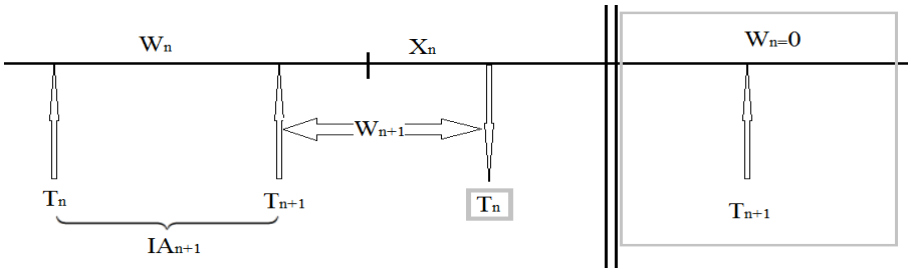

Fig. 5. Representation for calculating the waiting time, depending on the arrival of the (n+1)th customer

Recursively, when the next customer arrives, there are 2 possibilities:

- The arrival can occur after the n^{th}request was already serviced, therefore $\mathbf{W_{n+1}}$=0 (explained in the right grey-boxed part of figure 5), or
- The arrival occurs after $\mathbf{T_n}$ but before the n^{th} request leaves the system. From the fifth figure the waiting time of the $(n+1)^{th}$ request is deduced as the distance between its arrival and the moment when the n^{th}request leaves the system, mathematically represented as $\mathbf{W_{n+1}=W_n+X_n-IA_{n+1}}$, where $\mathbf{IA_{n+1}}$is the inter-arrival time between the n^{th} and $(n+1)^{th}$request.This can be easily translated into a single instruction that can be solved recursively using any modern programming language.

3.2. Performance measurements for the M/G/1 queue

If λ is the arrival rate and X is the service time, the server utilization is given by:

$$\rho = \begin{cases} \lambda E(X), if\ \lambda E(X) < 1 \\ 1, otherwise \end{cases} \quad (6)$$

Moreover, if the arrivals are described by a Poisson process, the probability that a request must wait in a queue is $P(W > 0) = \rho(7)$, and the mean waiting time is given by the Pollaczek-Khintchin formula[3] :

$$E(W) = \frac{\rho}{1-\rho} * \frac{E(X)}{2}(1 + \frac{V(X)}{E^2(X)}) \quad (8)$$

In addition, if the service times are exponentially distributed and the service follows the FIFO principle ("first-in-first-out", also knows as FCFS, "first-come-first-serve"), then the distribution function

of the waiting time is given by the following formula[1]:

$$F_W(t) = \begin{cases} 0, t < 0 \\ 1 - \rho e^{-(1-p)\frac{t}{E(X)}}, t \geq 0 \end{cases} \quad (9)$$

There is no simple formula for $F_w(t)$when the service times are not exponentially distributed, but using computer simulation can help developing such models, after validating classic models as the one above.

4.1. Software simulation of the Queueing Problem

As described previously, modeling the M/G/1 queue can be done by using a recursive algorithm by generating the inter-arrival time and the service times using the Inverse Transform Method[4].
The following lines written in the BASIC programming language simulate such an algorithm, although almost any programming language could be used.

```
100 FOR I=1 to 10000
110 IA= 'inter-arrival times to be generated
120 T=T+IA  'time of the next arrival
130 W=W+X-IA 'recursive calculation of waiting times
140 IF W<0 THEN W=0
150 IF W>0 THEN C=C+1 'count all requests that wait
160 SW=SW+W 'sum of waiting times for calculating E(W)
170 X= 'service times to  be generated
180 SX=SX+X 'sum of service times for calculating Utilization
190 NEXT I
200 PRINT SX / T, C / 10000, SW / 10000 ' print Utilization, P(W) and
E(W)
```

4.2. Generating random service and inter-arrival times using the Inverse Transform Method

Assuming that the computer can generate independent identically distributed values that are uniformly distributed in the interval (0,1), a proper method of generating random variable values according to any specified distribution function is using the Inverse Transform Method.
To generate the random number X, it is enough to input the randomcomputer generated number on the vertical axis andto project the value over the distribution function G, where G is the desired distribution to be generated. Projecting the point from the G graph furtherdown on the horizontal axis, delivers the desired randomly distributed values described by the G density function.This method is practically reduced to finding the inverse function of the distribution function of the distribution according to which the numbers are generated.By plugging in the computer randomly generated numbers, a new random variable is generated with has its distribution function G(u) [4]. This procedure is schematically described in the below figure.

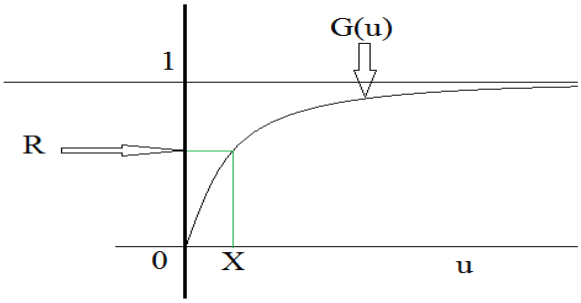

Fig.6. Illustration of the Inverse Transform Method

For example, for a Poisson process of arrivals that are exponentially distributed with parameter λ, where λ is the arrival rate and $\frac{1}{\lambda} = E(IA)$, according to the Inverse Transform Method, a value of λ=1.6 arrivals per second is derived, equivalent to anaverage inter-arrival time of $\frac{1}{\lambda} = \frac{5}{8}$seconds. For $G(u) = 1 - \lambda e^{-\lambda u} = R$ with u≥0, it is deduced that $G^{-1}(R) = -1/\lambda \ln(1-R)$ where R is the computer generated value. Therefore, the instruction 110 from section 3 of this paper becomes:110 IA=-(5/8)*LOG(1-RND), where RND is the BASIC function that generate values uniformly distributed between 0 and 1.Of course, any programming language that is able to generate random independent identically distributed numbersbetween 0 and 1 can be used for simulation.

5. Comparing the mathematical solution of the queueing problemwith the computer simulation

To illustrate the applicability of the software simulation, 4 different arrival times distributions are analyzed :

1. Exponential service time, with mean service time E(X)=0.5
2. Constant service time, X=0.5
3. Uniformly identical distributed service times between 0 and 1, X~U(0,1)
4. Service times of 1/3 have a probability of 90%,and service times of 2 have a probability of 10%.

For all 4 simulations, exponential distributed inter-arrivals with λ=1.6 are used as derived in 4.2 section. All calculations in the following table are done according to the formulas presented in section 3.2.

Table 1. Comparison between the mathematical and simulated results

		ρ		P(W>0)		P(W>0.5)		E(W)	
X	**Formula of X**	**Theory**	**Simulation**	**Theory**	**Simulation**	**Theory**	**Simulation**	**Theory**	**Simulation**
1	**-0.5*LOG(1-RND)**	**0.8**	**0.799436**	**0.8**	**0.799817**	**0.65498**	**0.654924**	**2**	**1.991853**
2	**0.5**	**0.8**	**0.799724**	**0.8**	**0.799895**	**NA**	**0.55622**	**1**	**0.997296**
3	**RND**	**0.8**	**0.800048**	**0.8**	**0.800103**	**NA**	**0.622625**	**1,(3)**	**1.332808**
4	**q = RND: IF q <= 0.9 THEN X = 1 / 3 ELSE X = 2**	**0.8**	**0.804667**	**0.8**	**0.799336**	**NA**	**0.616419**	**2**	**1.999094**

All 4 simulations have been chosen in such way that E(X)=0.5, and the distinction is done by choosing the service times with different distributions. Since the utilization is directly dependent on the arrival rate and mean arrival times, it is equal with 80% in all 4 cases. According to (7), the probability of waiting is also equal to 80% in all 4 cases.

In this simulation, the mean waiting time, as deduced from the Pollaczek-Khintchin(8) formula, confirms the accuracy of the simulation model, and gives insights also for the other cases, offering a clear approximation of the behavior of the designed system. It is interesting to observe that mean waiting time when having exponential service times is double in comparison with the mean waiting time when having constant service times, although the mean service time, the utilization and the probability of waiting are equal in both cases.

6. Conclusions

Based on all information presented in this paper, I can conclude that computer simulation is an important tool for the analysis of queues whose service times have any arbitrary specified distribution. In addition, the theoretical results for the special case of exponential service times(8) are extremely important because they can be used to check the logic and accuracy of the simulation, before extending it to more complex situations.

Moreover, such a simulation gives insight on how such a queue would behave as a result of different service times. Further, I consider that it offers a methodology for looking into more complicated cases, when a mathematical approach cannot help.

References

[1] R. B. Cooper, *Introduction to Queueing Theory, Second Edition*. New York: North Holland, 1981, pp. 208-232.

[2]S. Ghahramani, *Fundamentals of Probability with Stochastic Processes, Third Edition*. Upper Saddle River, Pearson Prentice Hall 2005, pp.284-292.

[3] L. Lakatos , "*A note on the Pollaczek-Khinchin Formula*", Annales Univ. Sci. Budapest., Sect. Comp. 29 pp. 83-91, 2008.

[4]K. Sigman, "Inverse Transform Method". Available : http://www.columbia.edu/~ks20/4404-Sigman/4404-Notes-ITM.pdf [January 15, 2015].

[5] K. Sigman, "Exact Simulation of the stationary distribution of the FIFO M/G/c Queue", J. Appl. Spec. Vol. 48A, pp. 209-213, 2011, Available : http://www.columbia.edu/~ks20/papers/QUESTA-KS-Exact.pdf [January 20, 2015]

Florin-Catalin ENACHEgraduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2008. Starting 2010 he holds a MASTER degree in the field of Economic Informatics, in the area of "Maximum Availability Architecture". His main domains of interest are : Computer Sciences, Database Architecture and Cloud Performance Management. Since 2014 he is a PhD. Candidate at the Bucharest University of Economic Studies, focusing his research on Performance management in Cloud environments.