Contact

Calea Dorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: 1
E-mail: editor@dbjournal.ro

Contents:

Implementation of Cloud Computing into VoIP

Floriana GEREA
Economic Informatics Department, Academy of Economic Studies, Bucharest, Romania
floriana.gerea@gmail.com

This article defines Cloud Computing and highlights key concepts, the benefits of using virtualization, its weaknesses and ways of combining it with classical VoIP technologies applied to large scale businesses.
The analysis takes into consideration management strategies and resources for better customer orientation and risk management all for sustaining the Service Level Agreement (SLA). An important issue in cloud computing can be security and for this reason there are several security solution presented.
***Keywords**: Cloud computing, VoIP, Data virtualization, DoS*

1 Introduction

Present article focuses on the importance of virtualization in VoIP technologies and how the virtualization can improve the VoIP communication technologies extending this article's conclusions.
Progress of research efforts in a new technology is contingent on having a rigorous organization of its knowledge domain and a comprehensive understanding of all the relevant components of this technology and their relationships.
The National Institute of Standards and Technology (NIST) defines cloud computing as a model for convenient, on-demand network access to computing resources such as networks, servers, storage, applications, and services that can be quickly deployed and released with very little management by the cloud-provider [4].
Cloud computing is an emerging technology from which many different industries and individuals can greatly benefit. Cloud computing services certainly have the potential to benefit both providers and users. However, in order for cloud computing to be practical and reliable, many existing issues must be resolved.
The use of cloud computing is particularly appreciated to users because it is rather inexpensive and it is very convenient. Users can access data or use applications with only a personal computer and internet access.
Over the years virtualization had an important role in developing IT projects. However which are the virtualization concepts, when can we use it and are there any risks that can appear in such implementation? These questions will be answered in the following paragraphs.
Also challenges in architectural design and security represent the main task that should be analyzed in implementing Cloud architecture for VoIP.

2. Architectural Cloud Computing

Cloud computing is quickly becoming one of the most popular new idea that will supposedly reshape the information technology (IT) services landscape. According to The Economist in a 2008 article, it will have huge impacts on the information technology industry, and also profoundly change the way people use computers [2]. What exactly is cloud computing then, and how will it have such a big impact on people and the companies they work for? In order to define cloud computing, it is first necessary to explain what is referenced by the phrase "The Cloud".
Cloud computing is in many ways a conglomerate of several different computing technologies and concepts like grid computing, virtualization, Service oriented

Architecture (SOA) [3], peer-to-peer (P2P) computing [2].

To begin understanding cloud computing, it is necessary to examine it in abstraction layers.

Figure 2 illustrates the five layers that constitute cloud computing [4]. A particular layer is classified above another if that layer's services can be composed of services provided by the layer beneath it.

The bottom layer is the physical hardware, namely the cloud-provider owned servers and switches that serve as the cloud's backbone. The next layer consists of the cloud's software kernel. This layer acts as a bridge between the data processing performed in the cloud's hardware layer and the software infrastructure layer which operates the hardware.

The abstraction layer above the software kernel is called software infrastructure. This layer renders basic network resources to the two layers above it in order to facilitate new cloud software environments and applications that can be delivered to end-users in the form of IT services.

The services offered in the software infrastructure layer can be separated into three different subcategories: computational resources, data storage, and communication. [5]

Several current examples of clouds that offer flexible amounts of computational resources to its customers include the Amazon Elastic Compute Cloud (EC2) [9], Enomaly's Elastic Computing Platform (ECP) [10], and RESERVOIR architecture [6]. Computational resources, also called Infrastructure as a Service (IaaS), are available to cloud customers in the form of virtual machines (VMs). Voice over Internet Protocol (VoIP) telephones, instant messaging, and audio and video conferencing are all possible services which could be offered by CaaS in the future.

All of three software infrastructure subcomponents, cloud-customers can rent virtual server time (and thus their storage space and processing power) to host web and online gaming servers, to store data, or to provide any other service that the customer desires.

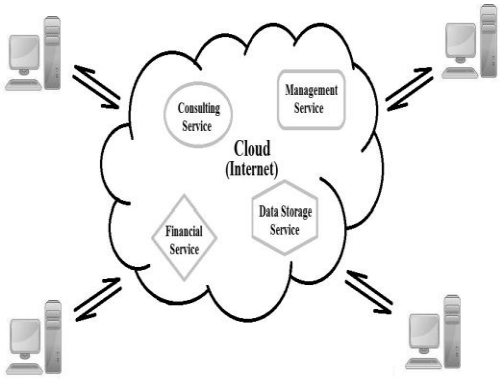


Fig. 1. Cloud computing [5]

When developers design their cloud software for a specific cloud environment, their applications are able to utilize dynamic scaling and load balancing as well as easily have access to other services provided by the cloud software environment provider like authentication and email. This makes designing a cloud application a much easier, faster, and more manageable task. There are several Virtual Infrastructure Managements in IaaS, such as CLEVER [25], Open-QRM [8], OpenNebula [6], and Nimbus [9].

Cloud management provides remote and secure interfaces for creating, controlling, and monitoring virtualized sources on an infrastructure-as-a-service cloud. VI management provides primitives to schedule and manage VMs across multiple physical hosts. VM managers provide simple primitives (start, stop, suspend) to manage VMs on a single host.

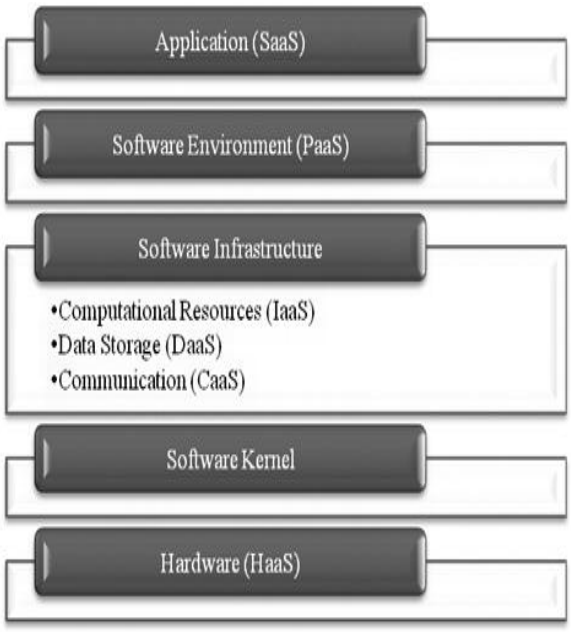


Fig. 2. The five abstraction layers of cloud computing [5]

3 Benefits from Cloud Computing

- *Access to Services that are otherwise unavailable.* In some circumstances, a service-provider may offer a new or exclusive capability - although this is likely to be the case only during a limited period of time, since most such services are, at least in principle, compatible. A more common situation is that some organizational users, and especially many individual users, may be technically or financially incapable of establishing and running a particular service for themselves
- *Access to Services from multiple desktop devices.* Each user, whether within an organization or acting as an individual, is likely to use multiple desktop devices, in various locations, including at home, at work, at clients' sites, in airport lounges, in Internet cafes, etc. By using authoritative data running on a remote device, the user reduces device-dependence in exchange for increased network-access dependence. In many circumstances, the trade-off may be advantageous
- Access to Services from scaled-down devices. A service may perform the vast majority of the function server-side, enabling the client to be run on a device with very limited capacity. This opens up scope for long-promised but little-used 'thin clients', but the primary devices used are more likely to be various forms of handheld computers, and mobile phones. This depends, however, on the service being designed with this aim in mind
- Access to Services from multiple device-types. Each user, whether within an organization or acting as an individual, is likely to use multiple kinds of devices, including desktop PCs, portable PCs, various forms of handheld computers, and mobile phones. A suitably-designed service may be able to support convenient access to data and applications on any and each of these device-types, through a variety of user-interfaces

Other Technical Benefits

- Professionalized backup and recovery. A service may be designed to provide assured backup of data and software, and assured, simple, efficient recovery. This is because these are core capabilities of a service provider, and that organization is likely to be more professional, attentive and disciplined than many user organizations and particularly individual users. Backup and recovery services can be provided whether the primary operational service is run on the user's own network, outsourced, or delegated to the cloud.
- *Scalability.* Where the transaction and/or data-volumes vary significantly over time, a service may offer assured server-capacity, storage-capacity, and access to the requisite application software. This may apply in a long-term growth curve (or indeed a tailing-off, as occurs with many legacy systems), and in contexts that involve highly-peaked demand, associated with daily, weekly, monthly, annual or even longer cycles, and with events

- *Collaboration convenience.* Collaborative content (including documents and other data which are co-owned and co-maintained) is inherently accessible and amendable by multiple authors. There are advantages in hosting a service such as a Wiki remotely from each of the participants. There may be advantages in the remote host being flexible rather than fixed
- *Copyright convenience.* The service-provider can assume responsibility for all aspects of acquisition, maintenance and licensing of software and of data

Few of the potential benefits arise solely from the incremental difference between cloud computing and its predecessors, and hence rational users need to consider whether cloud computing or some more conventional form of outsourcing, or indeed insourcing, is appropriate to their needs. Moreover, none of the benefits arise automatically, but rather are contingent on correspondence between the user's needs, on the one hand, and the service-provider's capabilities, terms of service and pricing, on the other.

Despite the technical benefits, it appears that service-providers perceive the primary driver for adoption as being cost-savings. A secondary driver may be convenience to business divisions arising from the ability to by-pass internal IT departments and contract directly for services. If this transpires to be the case, then the cautious risk assessment conventionally undertaken by IT departments will also be by-passed. It is therefore particularly important for senior executives to appreciate the downsides of cloud computing that are analyzed in the following sections. Technical factors are identified first, then business risks.

Despite all the benefits, there some security issues that are going to be discussed in the following part.

4 VoIP Cloud Computing vs. Traditional VoIP

In the traditional VoIP technology because the information is on a single server several problems can appear regarding data availability and integrity, security and in order to resolve these, money is spend on hosting software, applications and people with the requisite expertise. On the other hand Cloud Computing is less expensive because of its financial benefits.

Assuming that the hardware equipments can encounter several malfunctions, in a time when the services' quality is extremely important, the information needs to be available in real time. The traditional approach is to invest in a large number of equipments in order to avoid the loss of call and provide a correct functionality of the telephony service. However, these long term investments may be justified but at a closer analysis we can find that those equipments are not using all their resources. There has been statistically proven that most of the servers' hardware will never be fully used and as time passes they will be replaced due to moral and physical degrading.

Cloud computing can solve all these aspects. Organizations can avoid large investments in equipments and software by using a much smaller number of resources for one solution.

In this way investments can be made in fewer equipments with larger resources that are wiser employed, by creating a large number of virtual nods on one physical machine. By monitoring and controlling performance, organizations can easily decide which resources can be allocated on different services.

The reduction of operational costs

In cloud computing the organizations or the individual user are able to pay for only the services they need, avoiding the excess of employed resources that is involved in the traditional method.

Economy can be made, provided the service provider has an well-organized plan. In this way cloud computing has significant economical advantages comparing to the traditional method.

We also must mention the personnel costs that in the traditional method implies, because it requires a large number of people to mange resources, allocated in different geographical areas. Also, every new installation needs to be fully made, and this translates in large installation time for every new server. In cloud computing these aspects can be solved in a reduced amount of time, the installation of services taking very little. It is done by cloning other virtual nodes, so all the software and application installation is done only once and then all the new software is installed by cloning. In this way a large number of identical servers can be created within minutes, without the need to separately install each necessary application.

Cloud computing reduces human error to a minimum, due to the fact that there is no need to process the same information every time. It is enough to have only one correct virtual machine, that has been tested, all the other being replicas of the first.

- *Migrating services from one geographical area to another, from one machine to another, transferring from one solution to another*

The classical method required for each modification to restart all the installation procedures, which involved time spent and large costs. Cloud computing has the extraordinary benefit of easily moving information from one machine to another and between servers, without taking into account the geographical distance. It is possible for a virtual machine to have a node in Bucharest and to move that service within minutes on another server in Brasov, without damages or problems. Within minutes servers can be moved from one location to another, from one country to another, while keeping the service functional even while migrating.

This option did not exist in the traditional method. Using this method implied that the service would not be functional for at least several days, and that the physical movement of the server from one location to another was needed as well as a list of modifications that are necessary for any physical movement.

5 The levels that can attack a VoIP infrastructure

Denial-of-Service or VoIP Service Disruption. Denial-of-service (DoS) attacks can affect any IP-based network service. The impact of a DoS attack can range from mild service degradation to complete loss of service. There are several classes of DoS attacks. One type of attack in which packets can simply be flooded into or at the target network from multiple external sources is called a distributed denial-of-service (DDoS) attack.[3] DoS attacks are difficult to defend against, and because VoIP is just another IP network service, it is just as susceptible to DoS attack as any other IP network services. Additionally, DoS attacks are particularly effective against services such as VoIP and other real-time services, because these services are most sensitive to adverse network status. Viruses and worms are included in this category as they often cause DoS or DDoS due to the increased network traffic that they generate as part of their efforts to replicate and propagate.[9]

ARP Spoofing

ARP is a fundamental Ethernet protocol [3]. Perhaps for this reason, manipulation of ARP packets is a potent and frequent attack mechanism on VoIP networks. Most network administrators assume that deploying a fully switched network to the desktop prevents the ability of network users to sniff network traffic and potentially capture sensitive information traversing the network. Unfortunately, several techniques and tools exist that allow any user to sniff traffic on a switched network because ARP has no provision for authenticating queries or query replies [4].

Additionally, because ARP is a stateless protocol, most operating systems (Solaris is an exception) update their cache when receiving ARP reply, regardless of whether they have sent out an actual request.

H.323-Specific Attacks

The only existing vulnerabilities that we are aware of at this time take advantage of ASN.1 parsing defects in the first phase of H.225 data exchange. More vulnerability can be expected for several reasons: the large number of differing vendor implementations, the complex nature of this collection of protocols, problems with the various implementations of ASN.1/PER encoding/decoding, and the fact that these protocols —alone and in concert — have not endured the same level of scrutiny that other, more common protocols have been subjected to. For example, we have unpublished data that shows that flooding a gateway or media server with GRQ request packets (RAS registration request packets) results in a DoS against certain vendor gateway implementations—basically the phones deregister [9].

SIP-Specific Attacks

Multiple vendors have confirmed vulnerabilities in their respective SIP (Session Initiation Protocol) implementations [3].The vulnerabilities have been identified in the INVITE message used by two SIP endpoints during the initial call setup. The impact of successful exploitation of the vulnerabilities has not been disclosed but potentially could result in a compromise of a vulnerable device. In addition, many recent examples of SIP Denial of Service attacks have been reported.

Recent issues that affect Cisco SIP Proxy Server (SPS) demonstrate the problems SIP implementers may experience due to the highly modular architecture or this protocol. The SSL implementation in SPS (used to secure SIP sessions) is vulnerable to an ASN.1 BER decoding error similar to the one described for H.323 and other protocols. This example illustrates a general concern with SIP: As the SIP protocol links existing protocols and services together, all the classic vulnerabilities in services such as SSL, HTTP, and SMTP may resurface in the VOIP environment.

Policies and Processes

Encryption

All VoIP systems should use a form of Media (RTP channel) Encryption in order to avoid the sniffing of VoIP data. All communications between network elements should be encrypted. Complete end-to-end IP voice encryption is recommended to mitigate the threat of eavesdropping attempts. Additionally, all administrative access to critical server and network components must use encrypted protocols such as SSL and/or SSH [5]. All access to remote administrative functions should be restricted to connections to the switch itself or to a designated management PC [9].

Physical Security

Physical security is an essential part of any security plan [6]. Physical security refers to the protection of building sites and equipment (and all other information and software contained therein) from theft, intrusion, vandalism, natural disaster, man-made catastrophes, and accidental damage (e.g., from electrical surges, extreme temperatures, and spilled coffee). It requires suitable emergency preparedness, reliable power supplies, adequate climate control, and appropriate protection from intruders.

Safeguards can be broken down into two categories: human and environmental.

Human safeguard recommendations are:

- Console access should be restricted or eliminated.
- Logon, boot loader, and other passwords must be a minimum of eight characters including at least one each of alpha, numeric, and ctl characters.
- VoIP components must be located in a secure location that is locked and restricted to authorized personnel only.
- Access to these components, wiring, displays, and networks must be controlled by rules of least privilege.
- System configurations (i.e., hardware, wiring, displays, networks) must be documented. Installations and changes to those physical configurations must be

governed by a formal change management process.
- A system of monitoring and auditing physical access to VoIP components, wiring, displays, and networks must be implemented (e.g., badges, cameras, access logs). From the point at which an employee enters the building, it is recommended that there be a digital record of their presence.
- The server room should be arranged in a way that people outside the room cannot see the keyboard (thus seeing users/admin passwords).
- Any unused modems must be disabled/removed.
- No password evidence (notes, sticky notes, etc.) is allowed around the system.
- The CPU case should be locked and the key must be accounted for and protected. A backup key should be made and kept securely offsite (e.g., in a safety deposit box).
- USB, CD-ROM, monitor port, and floppy disks drives should be removed, disabled, or glued shut.
- Adequate temperature and humidity controls must be implemented to avoid equipment damage.
- Adequate surge protectors and UPS must be implemented, maintained, and tested.
- Cleaning and maintenance people should be prohibited from the area surrounding
any electronics.
- Food, drink, or smoking is prohibited in the same areas.

IP-PBX equipment must be located in a locked room with limited access. This type of access must be provided as a user authentication system with either a key-card or biometric device. The use of a keypad alone to gain access is not permitted. All methods of gaining entry into the room must provide for a list of users that have accessed the room along with a date/time-stamp [6].

6 Security for the VoIP Infrastructure

One example of how to configure a secure an system cloud for VoIP is the creation of a network demilitarized zone (DMZ) on a single host.

In this example, three virtual machines are configured to create a virtual DMZ on Standard Switch 1: Virtual Machine 1, 2,3 and 4 run Web server and are connected to virtual adapters through standard switches.

These virtual machines are multi homed. The Machine 5 and 6 runs an Asterisk server. The conduit between these elements is Standard Switch 2, which connects the firewalls with the servers. This switch has no direct connection with any elements outside. From an operational viewpoint, external traffic from the Internet enters Virtual Machine 1 through Hardware Network Adapter 1 (routed by Standard Switch 1) and is verified by the firewall installed on this machine. If the firewall authorizes the traffic, it is routed to the standard switch in the DMZ, Standard Switch 2. Because the Web server and application server are also connected to this switch, they can serve external requests. Standard Switch 2 is also connected to Virtual Machine 4 and Virtual Machine 5. This virtual machine provides a firewall between the DMZ and the internal corporate network.

This firewall filters packets from the Web server and application server. If a packet is verified, it is routed to Hardware Network Adapter 2 through Standard Switch 3. Hardware Network Adapter 2 is connected to the internal corporate network. This network could be used for virus propagation or targeted for other types of attacks. The security of the virtual machines in the DMZ is equivalent to separate physical machines connected to the same network.

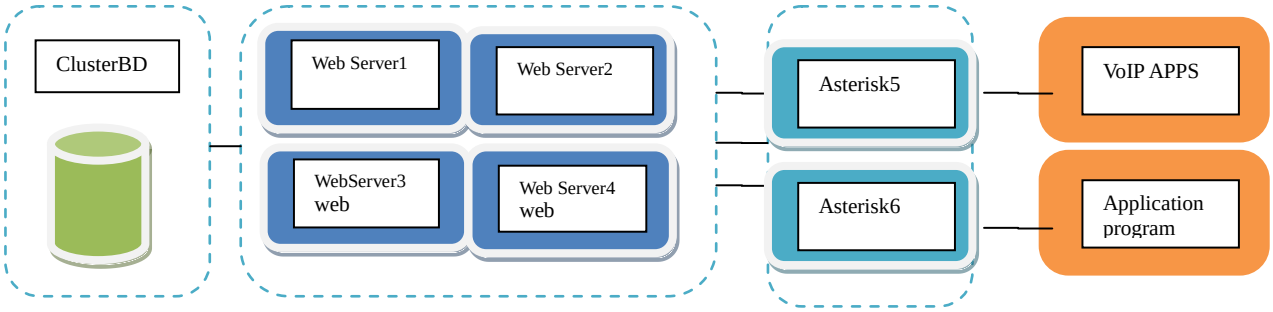


Fig. 3. Architecture VoIP

VoIP is a highly critical data application and as such, is subject to all the policies detailed in other data security policy sections (this assumes that the VoIP Security Policy module is part of a larger set of security policy modules).

Because in the cloud-based computing environment, the employees can easily access, falsify and divulge the data. Sometime such behaviour is a disaster for a big and famous company.

Some service providers develop some technical method aimed to avoid the security treats from the interior. For instance, some providers limit the authority to access and manage the hardware, monitor the procedures, and minimize the number of staff who has privilege to access the vital parts of the infrastructure. However, at the provider backend, the administrator can also access the customer's VM-machine.

Security within cloud computing is an especially worrisome issue because of the fact that the devices used to provide services do not belong to the users themselves. The users have no control of, nor any knowledge of, what could happen to their data. This, however, is becoming increasingly challenging because as security developments are made, there always seems to be someone to figure out a way to disable the security and take advantage of user information.

Traditional telephony, based on dedicated transmission lines, used over the last decades, has found through VoIP an important competitor, mainly because of the technology differences between them. In traditional telephony, through the switch-circuit system (circuit commuting), a communication channel between the two correspondents is assured. This channel (physic electric circuit obtained by cables and electronic circuits) must be assured before the communication starts. During the conversation the channel must be used only by the same initial correspondents, being a channel dedicated to communication. At the end of the conversation, this channel must be cancelled. This system was later improved by multiplexing of more channels on the same physic conductor, but each of these channels is dedicated only to one call at a certain time. In telecommunication, circuit commuting represents o routing method of the transmission between two correspondents, through one or more commuting centres. Between these two correspondents a continuous electronic connection is established, which will have the audio signal. The total of these telephonic central systems and of the connections that forms between them is called public network of commuting telephony (PSTN: Public Switched Telephone Network).

Improved functionality: another important advantage is that of a improved functionality as compared to classic telephony. Some of the functionalities offered by VoIP are difficult or even impossible to accomplish in the classic telephony. Among these, there is the possibility to use an IP telephone wherever there is a connexion to Internet. This creates the possibility that the "fix" telephone be taken in travelling, having the call number everywhere. The most important beneficiaries of this facility are the Call Centre agencies, that use VoIP

telephony in foreign countries due to the reduced costs with cheaper work force.

7 Conclusion

By innovation and a perfectible degree of security, VoIP industry is consolidating is market place, frightening to be able soon to take the place of conventional solutions (expensive, insecure and inflexible).

Cloud computing allows to create inexpensive systems, with little upfront costs and to be scaled to massive sizes, when needed. In many cases the best VOIP solution is to use cloud computing and replace the classical solution. The advantages can be defined both by the providers, which are motivated by the future profits that can arise due to the lower costs that the classical technology, as well as the users who have the possibility of reducing or eliminating the telephony service costs.

References

[1] G. Gruman, E. Knorr, What cloud computing really means. *InfoWorld*, (2009, May). [Online]. Available: http://www.infoworld.com/d/cloudcomputing/what-cloud-computing-reallymeans-031

[2] L. Siegele, Let it rise: A survey of corporate IT. *The Economist*, (Oct., 2008).

[3] P. Watson, P. Lord, F. Gibson, Panayiotis Periorellis, and Georgios Pitsilis. *Cloud computing for e-science with carmen*, (2008), pp. 1–5.

[4] R. M. Savola, A. Juhola, I. Uusitalo, Towards wider cloud service applicability by security, privacy and trust measurements. *International Conference on Application of Information and Communication Technologies (AICT)*, (Oct., 2010), pp. 1–6.

[5] M.-E. Begin, An egee comparative study: Grids and clouds – evolution or revolution. *EGEE III project Report*, vol. 30 (2008).

[6] B. Rochwerger, D. Breitgand, E. Levy, A. Galis, K. Nagin, I. M. Llorente, R. Montero, Y. Wolfsthal, E. Elmroth, J. Caceres, M.Benyehuda, W. Emmerich, F. Galan, The Reservoir model and architecture for open federated cloud computing. *IBM Journal of Research and Development*, vol. 53, no. 4 (July, 2009), pp. 1–11.

[8]"Implementing QoS Solutions for H.323 Videoconferencing over IP", *Cisco Systems Technical Whitepaper Document Id: 21662*, 2007.

[9] P. Calyam, M. Haffner, E. Ekici, C.-G. Lee, "Measuring Interaction QoE in Internet Videoconferencing", *Proc. of IFIP/IEEE MMNS*, 2007.

[10] S. Winkler, "Digital Video Quality: Vision Models and Metrics", *John Wiley and Sons Publication*, 2005.

Floriana GEREA is Security Analyst at Raiffeisen Bank. She has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2007. Currently, she is a PhD student in the field of Economic Informatics at the Academy of Economic Studies. She is co-author of one book ("Telecomunicatii si Tehnologia bazelor de date"), 6

published articles, and 2 scientific papers. Her fields of interest include: Linux, Clusters, VoIP and Cloud Computing.

Banking Intelligence Accelerator – Decision Support

Adrian MUNTEANU, Ovidiu RĂDUŢĂ
Economic Informatics Department, Academy of Economic Studies, Bucharest, ROMANIA
adrianm21@yahoo.com, ovidiu.raduta@gmail.com

Along with the development of information technology, Business Intelligence plays an important role in banking operation process. Business Intelligence in banking sector is a method of storing and presenting key bank business data so that any key user can quickly and easily ask questions of accurate and timely data. The growing competition and increased speed of business changes has dramatically shown the need for business intelligence in banking sector. In this paper, we analyze the business intelligence components, how they fit the banking sector and how they can be secured to match the framework of the whole banking information system. Having the decision process analyzed in the banking field, we propose an architectural model to sustain the decision and integrate easily in the complex banking environment.

Keywords: *Business Intelligence, Decision, Data Warehouse, Architecture, Banking Accelerator*

1 Introduction

The world wide emergence of information revolution impacts every type of business and industry, and particularly, the banking industry. The value of needed piece of information increases with the square root of the number of users who can access that information, multiplied by the number of business areas in which users act. Inspite the huge amount of information stored inside banking information systems about customers and their transactions, the banks can rarely exploit its full potential in leveraging of tactical and strategic decision making [6]. Striving for a success, banks are trying to find means for efficient analysis of these data. Implementation of BI systems in banking begins with collection, enhancement, and purification of the daily legacy data from internal and external sources, including third party organizations. Availability of "enhanced" data in real time helps banks recognize and take benefit from new possibilities to strengthen customer relations, attract new prospects, and adapt to growth. BI effectively couples business strategies with information technologies leveraging on the existing IT infrastructure and skills [5].

2 Banking Intelligence Overview

Industry experts are in agreement that today's banking industry enjoys a highly dynamic environment – with mergers, acquisitions, divestitures, outsourcing initiatives, and branches closing and reopening – characterized by a fast-changing set of business, regulatory, and IT requirements. It is imperative for banks to gain market share while increasing operational efficiency. An interesting view from Forrester Research is that on the IT side, existing banking platforms can be compared with baroque castles. These ancient constructs are not up to the quality, cost and time-to-delivery requirements of modern banks.

The definition we agreed for the Banking Intelligence was the ability of a banking organization to comprehend and use information in order to increase its key performance indicators. Banking Intelligence comprises of a number of activities, procedures and applications, some of mostly used are: Data Warehousing, Data Marts, OLAP tools, tools for Extraction, Transformation, and Loading (ETL) of data, Information Portals, Data Mining, Business Modeling, etc. In this paper, we briefly describe three,

most commonly adopted technologies: Data Warehousing, Analytical processing, and Data Mining [9]. The banking sector has constantly been pushed by demands for new and innovative products and by regulatory requirements. Undergoing processes of a bank influenced by economy actions happening around the globe have inevitable made bank's information systems highly heterogeneous, with disintegrated applications, overlapping sets of data, and disperse points (in location and time) of data collection and processing. The idea to collect and unify the data form disparate sources has led to the concept of Data Warehousing. Data Warehouse filled with complete and purified (cleansed and enhanced) data is a prerequisite for the task of transforming information into knowledge. On-Line Analytical Processing and Data Mining are common methods for retrieving hidden knowledge from the data stored in a Data Warehouse. To picture the aforementioned Banking Intelligence components and emphasis how they fit into the whole banking system architecture, we illustrate **figure 1**, below, briefly resuming the main BI components and how the data flows between them, providing desidered output.

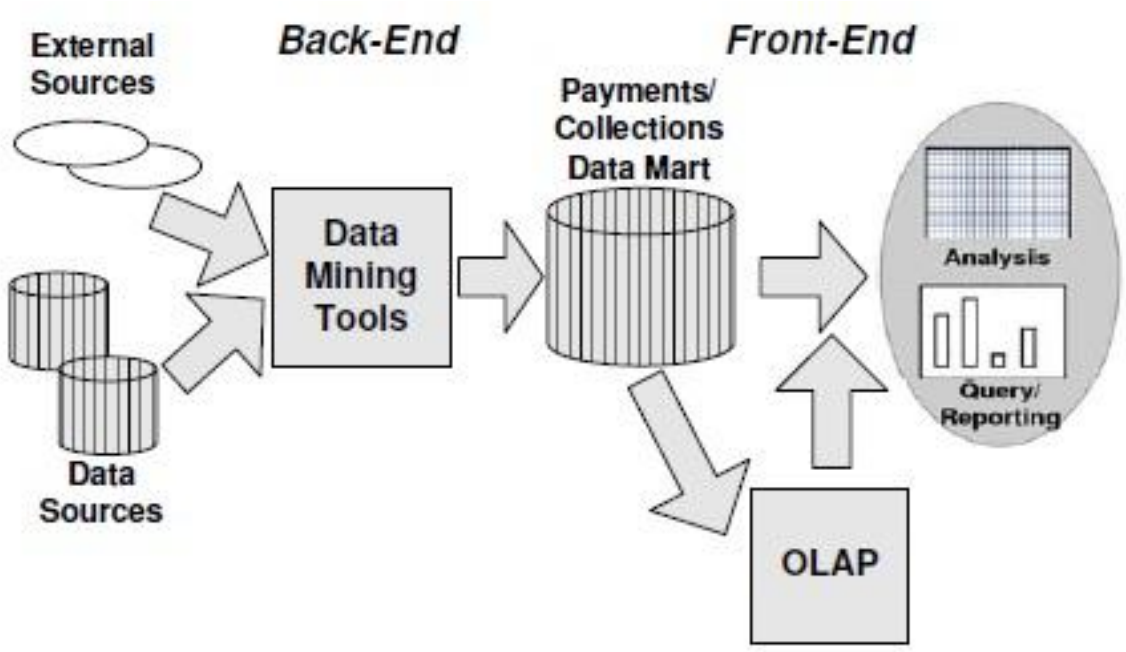


Fig. 1. Banking Intelligence architecture [3].

The picture depicts the core of the system comprised as a centralized data warehouse, allowing the end user to perform predictive, multidimensional analysis, over millions of transactions from customers of a bank agency. Also we have identified two innovative preprocessing methodologies, based on data mining techniques, which are exploited in the banking sector to populate the warehouse. These are highly effective at dealing with schema reconciliation and duplicate detection issues and, hence, assure a reliable integration and consolidation of heterogeneous bank data into a unique archive.

On-line Analytical Processing (OLAP) enables manipulation and analysis of large amount of data, comparison of different types of data, complex computations and, most importantly, an intuitive graphical user interface (GUI) for presentation of results in various perspectives including drill-up and drill-down capabilities. OLAP tools are essential component of today's BI systems [5]. Data Mining is founded on algorithms for detection of unknown and unexpected patterns in large sets of data,

clustering and segmenting of data and finding dependencies between multidimensional variables. The results of Data Mining analysis are presented graphically with the dominant and unexpected behavioral patterns enhanced. Applications for Data Warehousing, On-line Analytical Processing, and Data Mining are being widely adopted in modern banks to provide timely answers to many questions which previously required costly and lengthy programming and batch processing [7].

Systems of a Banking Intelligence standard, combine data from internal information systems of a bank and they integrate data coming from the environment e.g. statistics, financial and investment portals and miscellaneous databases. They are meant to provide adequate and reliable up-to-date information on different aspects of enterprise activities, having as final goal supporting decision in real time. The structure of such system consists of the following modules:

• tools to extract and transfer data – they are mainly responsible for data transfer from transactional systems and Internet to data warehouses;

• data warehouses – they provide room for thematic storing of aggregated and already analyzed data;

• analytic tools (OLAP) – they let users access, analyze and model business problems and share information stored in data warehouses;

• tools for reporting and ad hoc inquiring – they enable creation and usage of different synthetic reports;

• presentation layer – applications including graphic and multimedia interfaces whose task is to provide users with information in a comfortable and accessible form

3 Banking Intelligence key role in decision process

Business Intelligence is currently one of the fastest developing directions in information technology. Nowadays BI systems are connected with CRM systems (Customer Relationships Management) and ERP (Enterprise Resource Planning) to provide an enterprise with a huge competitive advantage[2]. Banking executives are focused on results. They need to know the bank's customers in great depth so as to sell the right product to the right customer for the right reasons and also manage risk and comply with changing requirements. All these challenges lead to the introduction of banking intelligence, which refers to technologies, applications and practices for the collection, integration, analysis, and presentation of business information and also sometimes to the information itself. Banking intelligence can provide historical, current, and predictive views of banking operations, most often using data that has been gathered into a data warehouse or a data mart. Main tasks that are to be faced by the Banking Intelligence systems include intelligent exploration, integration, aggregation and a multidimensional analysis of data originating from various information resources [8].

Based on our experience we could analyze and summarize main characteristics that a business intelligence application should meet in order to serve the decision support, its ultimate goal:

- To have a clear architecture. The BI solution employs the prevalent data warehouse concept. Data are collected into the data warehouse first, and then transferred into a specially designed data mart. Many banks have their own data warehouse. Aside of those, various data marts can be built for various purposes. The Banking Intelligence solution aims to adapting the existed warehouse and can be quickly deployed using this architecture.
- To be subject oriented. Based on our experience of banking industry, especially on the domestic market, we summarize an overall subject library. Those subjects are organized as main topic,

sub-topic, etc. The bank decidents, as their need, can select any subject they are interested in. The subject oriented manner makes the banking intelligence implementation unambiguous and cost effective.

- Prove timely response. This benefit inherits from the OLAP technology. Other than the off-line analysis and report, banking intelligence can trace the ongoing changes and provide valuable information by timely response.

The design of the data mart plays a key role in the process of decision making. In particular, the identification of its dimensions actually determines the capability of the overall decision support system to answer meaningful business queries [4]. However, many difficulties arise while constructing the foresaid warehouse. More specifically, these divide into syntactic and semantic issues. We could distinguish this involves two major problems, namely schema reconciliation and data reconciliation. In our banking intelligence scenario, the requirement for schema reconciliation mainly follows from the textual format of the personal information concerning both debtors and creditors. This information may not be uniformly formatted and, hence, its constituting records may apparently conform to different schemas. Indeed, the order of appearance of personal information attributes across the individual lines of text may not be fixed. In addition, their recognition is further complicated by the absence of both a canonical encoding format and of suitable field separators, which is mainly due to erroneous dataentry, misspelled terms, transposition oversights, inconsistent data collection and so forth. Also, distinct records may lack different attribute values, which makes them appear with a variable structure. Yet, the collection of personal information may be fragmented over disparate data sources, which further exacerbates the aforementioned difficulties.

4 Banking Intelligence Accelerator Model

Industry trends indicate that increasing competition in the financial industry is driving financial institutions to seek a customer - centric view of their business, in order to better understand their customers, deepen the business relationship with them, up-sell more products, and reduce turnover. In addition, consolidation and acquisitions within the financial industry are placing increased pressure on the ability of financial institutions to achieve that single view of the client, maintain consistent and reliable data quality, and control their exposure to regulatory and compliance risk.

What we propose in this paper is an accelerator that enables financial institutions to adopt a customer-centric view of performance by consolidating customer information across all lines of business. This accelerator solution incorporates each component of the transformation process to a customer-centric view. At its core, the accelerator is based on a time-tested Banking Data Model coupled with tools that help populate the model with the institution’s data. With this foundation in place, financial institutions would be able to leverage advanced analytics, predictive modeling, pre-built dashboards, and custom reports, to allow management at the branch, line of business, call center and enterprise levels to visualize performance and monitor customer behavior, thereby helping to improve and expand customer relationships.

To illustrate the Banking accelerator proposed solution we have put together **figure 2**, which shows a high components overview of proposed model.

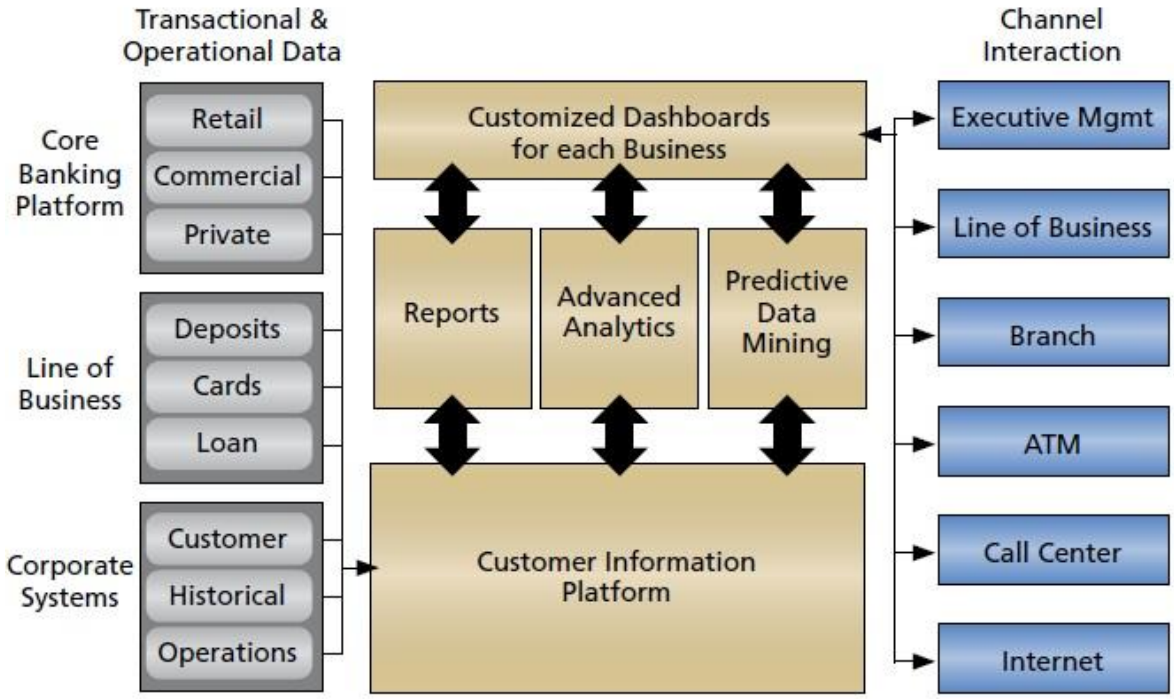


Fig. 2. Banking Intelligence Accelerator Model.

Implementing Banking accelerator solution, will transform the way Bank organizations run their business by providing them with pre-configured business products and ready-to-use transaction sets. With model proposed, banks can get a head start in core Banking deployment and based on a parameterization capability, banks can launch products quickly, and gain edge over competition and thereby significantly reduce the deployment risks, efforts and timeframe. The enterprise integration accelerators allow banks to easily migrate from their legacy core banking applications and in turn support them in their transformation initiatives. With reduction in deployment time, banks can focus on their growth objectives and deliver innovative products and services to customers.

Banks can leverage the proposed accelerator model to quickly expand into new customer segments and/or geographies. Using an open architecture, the model, could offer the option of implementing modules based on business relevance and should also allow the banks to leverage their current IT investments with its ability to co-exist with any third party application.

Such model can be built on new-generation technology platforms and should be capable of addressing multi-entity, multi-currency, multi-lingual, requirements. The Banking Intelligence Accelerator model can provide complete scalability and adaptability to rapidly growing and changing businesses like the banking one. It can be presented with a complete range of independent business modules, which operate within an integrated framework. This would allow the bank or financial institution to choose the module set that is relevant to their business paradigm and also allow the institution to decide on the sequencing the roll out of specific business areas during implementation.

5 Example of Banking Intelligence Accelerator Architecture

With the above model specification in mind we can propose an example of architecture that is based on the n-tier principles. High performance is ensured by basing the product on an N-tier

architecture, using messaging, by developing the middleware (glue binding the various tiers) as highly optimized & tuned in-memory processes. Each of the tiers performs a well defined function, thus allowing for a clear separation on processing. Each tier is independently scalable and designed to utilize the underlying multi-threading design and multi-processor hardware capability.

Having its architecture based on open-source components, it can be implemented and integrated in any banking environment regardless the hosting platforms. This advantage provides the banks with the ability to undertake major growth initiatives without being restricted by the application's inability to support those initiatives due to compatibility restrictions.

An illustrated example of such architecture that can consist in an implemented solution which should match the N-tier architecture and the model we propose, is shown in **figure 3**.

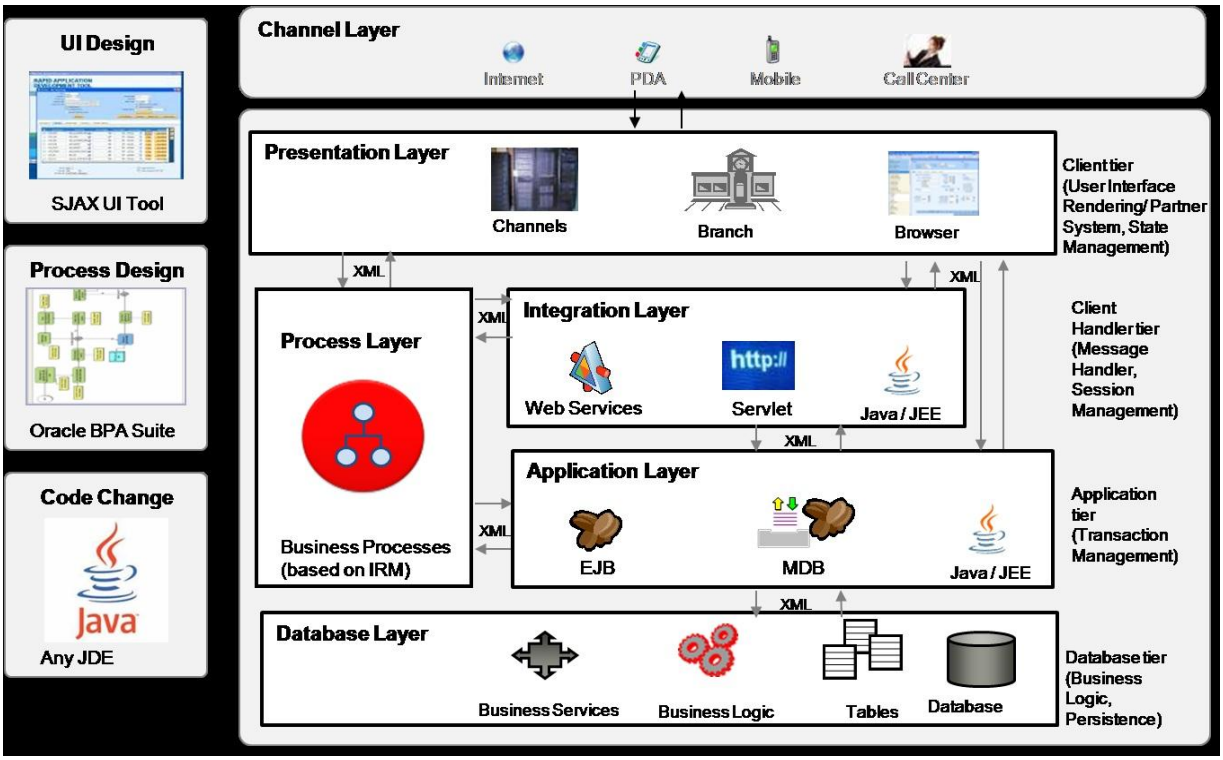


Fig. 3 Example of Banking Intelligence Accelerator architecture [10].

Reviewing this architecture, the system can consist of an *user interface*, preferably light-weight in nature (available on the web) and based on JavaScript and XML technologies. The communication between the browser and the web server can be using XML. The rendering is done on the client using XSLT for rendering as HTML. This allows for reduced infrastructure requirements due to following two most encountered restrictions: the bandwidth between the client and server is low and usually security policies might restrict a dedicated session with the database. The user interface should be configurable, while the screen can be easily adapted to different languages. Last, but not least, having in mind the Banking Intelligence Accelerator architecture is proposed for the banking sector, the user interface should meet the accessibility requirements for the teller (retail banking) screens and mobile devices.

The lower tier, we envisioned for the purpose of this paper, is the *process tier*, which should provide the ability for processes to be developed around the natively provided application. For example, can be defined processes using Oracle BPEL Process Manager and

integrate the same into the application‘s user interface framework. When deployed in a process centric model, the Banking Intelligence Accelerator provides a task-based user interface. By default, task based user interface is offered for the branch platform. The banking Intelligence accelerator interfaces with the Process Manager API‘s for viewing and managing tasks in the application. The process framework offers a facility to do the following: message content to control the data flow between different tiers, data transformation as needed, routines to match the banking business rules, flow control to monitor how the data is routed between different steps in the process and security to control who is eligible to perform a particular task.

At this level, the proposed architecture would not differentiate partner channels from its own native user interface when it comes to data processing. The *application and integration tier* provides the message handling, session management (for the native user interface) and transaction management in the application. The application and integration tier is based on open-source protocols like the ones offered by Java and JEE 5 technologies and is designed to be vendor independent, which makes the solution supported on any J2SE 1.4/JEE 5 compliant application server.

The back-end is a relational database management system i.e. Oracle 11g. The database tier ensures integrity of data and also provides business logic that can be PL/SQL stored procedures loaded into the database. At this tier, a clusterization solution i.e. Oracle RAC (Real Application Clusters) can be implemented to build redundancy and support failover & recovery, scalability and high availability.

We also envisioned the ability of extensibility framework for the Banking Intelligence Accelerator, which implies existence of additional handlers in the front-end units as well as back-end units. Typically, whenever any customization is done, the changes are typically done in the base units and if there are no additional handlers for site specific customizations like defaulting of specific values, inclusion of additional validations etc. the underlying base units need to be modified for any type of customization and this also makes identification of customization changes difficult.

Therefore, to minimize modification of base units, the architecture proposes at application tier level a provision to support extensibility in the following areas: maintenance of user defined fields at each screen level and capture of data, usage of user defined fields in advice formats, generation of notifications for operations done in the system.

It is a well recognized fact that any bank‘s data centre is bound to be heterogeneous in nature. Many systems, disparate technologies, various modes of usage, a multitude of mechanisms by which to integrate with, the Banking Intelligence accelerator needs to be fit into such an environment. The integration needs therefore, will also be varied in nature. Hence one of the design goals of proposed architecture has been that it should integrate easily with external systems for which it provides an *integration gateway*. *Integration gateway* uses a service and notifications based approach to integration under the principles of Service Oriented Architecture. The architecture provides for interfaces supporting open standards that can be used to easily integrate with other applications. The gateway provides for various deployment patterns, all of which can co-exist, to enable these various integrations.

6 Banking Intelligence accelerator – Implementation aspects

In the way our system’s architecture has been exposed in this paper, the implementation should meet out-of-the-box paradigm. This is because of its open-source and heterogeneous protocols and technology, the propose components connectors and handlers and the ability to

fit in banking and financial services business models.

System of the Banking Intelligence standard should be analyzed taking into consideration all the benefits of their implementation in a bank, likely to be generated [1]. Case studies we came through, proved and emphasised our belive that a Banking Intelligence solution, like the one presented hereby, may be utilized mainly for:

1. Strategic planning including first of all:
   - modeling different methods in the development of an organization;
   - informing about the realization of an bank's strategy, mission, goals and tasks;
   - identifying problems and "bottlenecks" to be tackled;
   - providing information on the enterprise's environment and market trends.
2. Improving relations with customers and in particular:
   - providing sales representatives with adequate knowledge on customers so that the reps could rapidly meet their customers' needs;
   - following the level of customer satisfaction, together with efficiency of business practices and identifying market trends.
3. Analyzing profitability of products and services manifested inter alia in:
   - providing analyses of "the best" and "the worst" products, employees, regions (as far as sales, costs or results are concerned).
4. Analyzing internal processes and operational efficiency of a banking system by means of:
   - providing analyses of deviations from the realization of plans;
   - providing knowledge and experience emerged while developing and launching new products onto the market;
   - exchange of knowledge among research teams and banking departments.
5. Controlling and management accounting and in particular:
   - analysis of actual costs and financial flows.

Due to solutions that fulfill all the functions mentioned above, management of a banking system gains new quality and, what is more an organization like banks is bound to become intelligent.

7 Conclusions

Scope of this paper was to identify the role of banking intelligence in decision processes, proposed a model to accelerate banking business processes, assisted decision, integrate easily in the multitude of banking environment applications and depict some of the most important aspects of implementation. The experience and researches in business intelligence and banking area, were a great help, understanding the technological and business processes running in the industry backend and build a model which should answer with ease the most recent requirements for decision process in financial services field.

Conclusion we came to, after conducted research for building this paper is that disconnected information systems can be integrated in order to achieve a complete and coherent decision support solution, so using business intelligence application standards is a primary rule for effective and documented decisions. In the banking sector, there is a spent of resources of annual IT budgets on business intelligence software products and they need to prove reliable and secure. Therefore, as we described across this paper, players in this area need a reliable infrastructure to link the explosion of new software applications and data sources, keeping a good pace with the volume of enterprise data, business partners, suppliers and customers.

8 Acknowledgment

The authors thank Ph.D. Manole Velicanu for the guidelines provided throughout the entire doctoral stages and his high level of co-operation for carrying out this study.

References

[1] C. Rayns, N. Ashworth, P. Beevers, V. Eibel, F. Jarassat, C. T. Jensen, A. Lucas, A. Papageorgiou, A. Roessle, N. Williams, "*Smarter Banking with CICS Transaction Server*", IBM Redbooks, pp. 26, 2010.

[2] D. Litan, L. Copcea (Teohari), M. Teohari, A.M. Mocanu (Virgolici), I. Surugiu, O. Raduta, "Information Systems Integration, a New Trend in Business", *APPLICATIONS of COMPUTER ENGINEERING (ACE '11)*, Canary Islands, ES, 2011.

[3] G. Costa, F. Folino, A. Locane, G. Manco, R. Ortale, "Data Mining for Effective Risk Analysis in a Bank Intelligence Scenario", *Data Engineering Workshop, IEEE 23rd International Conference*, pp. 2, 2007.

[4] J. Taylor, N. Raden, "*Smart (Enough) Systems: How to Deliver Competitive Advantage by Automating the Decisions Hidden in Your Business*", Pretince Hall, pp. 230, 2007.

[5] K. Curko, M. P. Bach, "Business Intelligence and Business Process Management in Banking Operations", *Information Technology Interfaces, The 29th International Conference*, pp. 2, 2007.

[6] Liautaud B., Hammond M., "*e-Business Intelligence, Turning Information into Knowledge into Profit*", McGraw-Hill, pp. 145-152, 2007.

[7] M. Velicanu, D. Litan, L. Copcea (Teohari), M. Teohari, A.M. Mocanu (Virgolici), I. Surugiu, O. Raduta, "Ways to Increase the Efficiency of Information Systems", *The Proc. of the 10th WSEAS Internat. Conf. on artificial Intelligence, Knowledge Engineering and Databases*, Cambridge, UK, pp.3, 2011.

[8] S. Mansmann, Neumuth, M. H. Scholl, "*An OLAP Technology for Business Process Intelligence: Challenges and Solutions*", Data Warehousing and Knowledge Discovery, pp. 143, 2007.

[9] Z. Lin, M. Zhu, W. Yin, J. Dong, "Banking Intelligence: Application of Data Warehouse in Bank Operations", *Service Operations and Logistics, and Informatics, IEEE/SOLI IEEE International Conference*, pp. 2, 2008.

[10] Mahmood Shah, Steve Clarke, E-Banking *Management: Issues, Solutions, and Strategies*, IGI Global, pp. 93, 2010.

Adrian Munteanu has graduated the Academy of Economic Studies (Bucharest, Romania), Cybernetics, Statistics and Economic Informatics in 2001. Currently, he is a Ph.D. Candidate in Economic Informatics with his Doctor's Degree Thesis: DataWarehouses - Business Support. In present, he is Advanced Resolution Engineer with 12+ years experience in database and Enterprise solutions field at Oracle Corporation. His research activity can be observed in many international proceedings (papers ISI proceedings) published by now. His scientific fields of interest include: Business Intelligence, Datawarehouse Modelling and Enterprise Resource Planning implementation.

Ovidiu Răduţă has graduated the Academy of Economic Studies (Bucharest, Romania), Faculty of Cybernetics, Statistics and Economic Informatics in 2006. He holds a Master diploma in Informatics Security (Master Thesis: IT Software in banks. Security Issues) from 2008 and currently, he is a Ph.D. Candidate in Economic Informatics with his Doctor's Degree Thesis: Bank System's Process Optimizing. In present, he is ISTQB – Advanced Test Analyst certified and he works as Senior Test Analyst with 3+ years testing

experience in Raiffeisen Bank Romania (6+ years banking projects experience). His research activity can be observed in many international proceedings (papers ISI proceedings). His scientific fields of interest include: Test management, Test Techniques, Databases processes, Middleware Products, Information Systems and Economics.

Clustering Analysis for Credit Default Probabilities in a Retail Bank Portfolio

Adela Ioana TUDOR, Adela BÂRA, Elena ANDREI (DRAGOMIR)
Bucharest Academy of Economic Studies
adela_sw@yahoo.com, bara.adela@ie.ase.ro, elena.andrei@gmail.com

Methods underlying cluster analysis are very useful in data analysis, especially when the processed volume of data is very large, so that it becomes impossible to extract essential information, unless specific instruments are used to summarize and structure the gross information. In this context, cluster analysis techniques are used particularly, for systematic information analysis. The aim of this article is to build an useful model for banking field, based on data mining techniques, by dividing the groups of borrowers into clusters, in order to obtain a profile of the customers (debtors and good payers). We assume that a class is appropriate if it contains members that have a high degree of similarity and the standard method for measuring the similarity within a group shows the lowest variance. After clustering, data mining techniques are implemented on the cluster with bad debtors, reaching a very high accuracy after implementation. The paper is structured as follows: Section 2 describes the model for data analysis based on a specific scoring model that we proposed. In section 3, we present a cluster analysis using K-means algorithm and the DM models are applied on a specific cluster. Section 4 shows the conclusions.
Keywords: *Data Mining, Cluster Analysis, Artificial Intelligence*

1 Introduction

Data mining is a technique that consists of analysing large volumes of information stored in data warehouses, in order to resolve decision problems. The technique is derived from three categories of software applications: the statistical ones, artificial intelligence applications based on neuro-fuzzy algorithms and the ones based on automated machine learning. Once the data has been prepared, the next step is to generate previously unknown patterns from the data using inductive learning. The most popular types of patterns are classification models, clustering and association rules that describe relations between attributes. Although methods and knowledge extraction techniques are applied in automatic mode, the process requires considerable human effort involved especially in the stages of analysis, but also in those of validation the results. There is a great deal of overlap between data mining and statistics. In fact most of the techniques used in data mining can be placed in a statistical framework.

The steps taken in the entire process are [6]:

- collect data from multiple sources: web, text, databases, data warehouses ;
- data filtering by eliminating errors. When using a data warehouse, this process is removed because a process of extraction, transformation and loading (ETL) was already applied on the data ;
- establishing key data attributes that will participate in the DM process, by selecting those properties that interest the analysis ;
- application of templates and detection / analysis of new knowledge ;
- visualization, validation and evaluation of results.

The steps in the mining process are performed iteratively until meaningful business knowledge is extracted.

The main idea of the article is to build a model based on data mining techniques that can predict customer behaviour over time. We used hierarchical clustering method on

the set of records, in order to obtain a profile of the customers. Then, we applied the data mining models on the cluster with bad debtors, reaching a very high accuracy after implementation.

The paper is structured as follows: Section 2 describes the model for data analysis based on a specific scoring model that we proposed. In section 3, we present a cluster analysis using K-means algorithm and the DM models are applied on a specific cluster. Section 4 shows the conclusions.

For testing different methods we used Oracle Data Mining (ODM) that is organized around several generic operations, providing an unified interface for extraction and discovery functions. These operations include functions for construction, implementation, testing and manipulation of data to create models. ODM implements a series of algorithms for classification, prediction, regression, clustering, association, selection, and data analysis. Oracle Data Miner provides the following options for each stage: for transforming the data and build models (build), for testing the results (test) and for the evaluation and application on new data sets (apply).

2 The model for data analysis

The study is based on financial data in 2009 from an important bank in Romania and the target customers are credit card holders. Among the 18239 instances, 1489 record arrears. The research involves normalization of attribute values and a binary variable as response variable: 1 for the default situation and 0 for non-default. Explanatory variables or the attributes are found in the scoring model proposed by us that includes: Credit amount (the amount limit by which the debtor may have multiple withdrawals and repayments from the credit card), Credit balance, Opening date of the credit account, credit product identification, Category, Currency, Client's name, Gender, Age, Marital status, Profession, Client with history (including other banking products and payment history), Deposit state, Amounts of deposits opened in the bank, if applicable, Scoring rate (scoring points from 1 to 6, 1 is for the best and 6 for the weakest debtor).

Data was divided into two tables, one used for model construction and one for testing and validating the model. Each case contains a set of attributes, of which one is the profiling attribute, this attribute is called 'RESTANTIER' that means that the creditor is a bad payer if the value is 1 and is a good payer if the value is 0.

First we applied three data mining models on the set of instances: classification, Naïve Bayes and regression with support vector machines. We obtained a series of predictions for credits reimbursement and the comparative results show that only 845 of 18 239 records are incorrect predictions, representing 4.63% of total.

We also made a comparison of the incorrect predictions released by the three models: SVM registered 406 incorrect predictions (2,22%), NB recorded 703 incorrect predictions (3,85%), and LG registered only 324 wrong predictions (1,77%). From cost point of view, LG recorded the lowest cost, followed by SVM, and NB is detaching pretty much. Although, we can consider that all three models can be successfully applied in banking practice, we extended our study with a clustering analysis. The following section presents the results.

3 Cluster analysis

3.1. Clusters building

Cluster analysis is a collection of statistical methods, which identifies groups of samples that behave similarly or show similar characteristics. In common parlance it is also called look-a-like groups. The simplest mechanism is to partition the samples using measurements that capture similarity or distance between samples [3].

Performing a cluster analysis targeting the classification of a set of objects, includes the following steps:

- choosing the characteristics subject to classification;
- choosing the measure in order to assess the proximity of objects;
- setting rules for grouping the classes or clusters;
- building the classes (ie classification of objects into classes);
- checking the consistency and significance of classification;
- choosing an optimal number of clusters, depending on the nature of the classification problem and the purpose;
- interpreting the significance of clusters.

We could say that cluster analysis can be understood as a classification technique or algorithm for organizing the data as classes or representative structures that verify certain properties. The results of cluster analysis are represented either by a single cluster solution, or cluster hierarchies, containing different ways of configuration of the objects in classes (ie cluster solutions).

We applied the clustering method on the set of records, taking into account only the important attributes. When building the model, we specified a number of 8 clusters, considering our objectives. Therefore, we took into consideration that the classified objects in each group to be as similar in terms of certain features and the classified objects into a group to differentiate as much as possible of the objects classified in any of the other groups.

The first criterion requires that each class to be as homogeneous compared to the characteristics considered for the classification of objects. The second criterion requires that each class may vary as much as possible in terms of classification features. A difficult problem that arises in cluster analysis is related to the need to assess the distances between classes or clusters.

K-means clustering is an iterative clustering method, and divides the data into a number of clusters by minimizing an error function which can be expressed. The *K-means* algorithm is a non-hierarchical approach to forming good clusters, used to group records based on similarity of values for a set of objects. Applying the concept allows classification for multiple classes and nonlinear relationships modelling between data (for prediction purposes). Even though, there may be difficulties often in establishing effective metrics, the technology being one of the few that accepts as input data of different nature (continuous, categorical, Boolean, etc.). Since the computing time is directly proportional to the number of instances in the data set, in the pre-processing stage, it is required to be selected from the original data set a subset of instances of reasonable size. This method proved to be effective in classification problems, when all associated key attribute classes have an equal representation as a percentage of the dataset. The algorithm based on k-NN technique allows only making an estimation of key attribute value, without generating additional information about the instances under review, the structure of the data set of classification categories of key attribute. This technique is used mainly in situations where for all attributes, the same function of distance is applied.

Figure 1 shows the cluster distribution and the number of observations in each cluster:

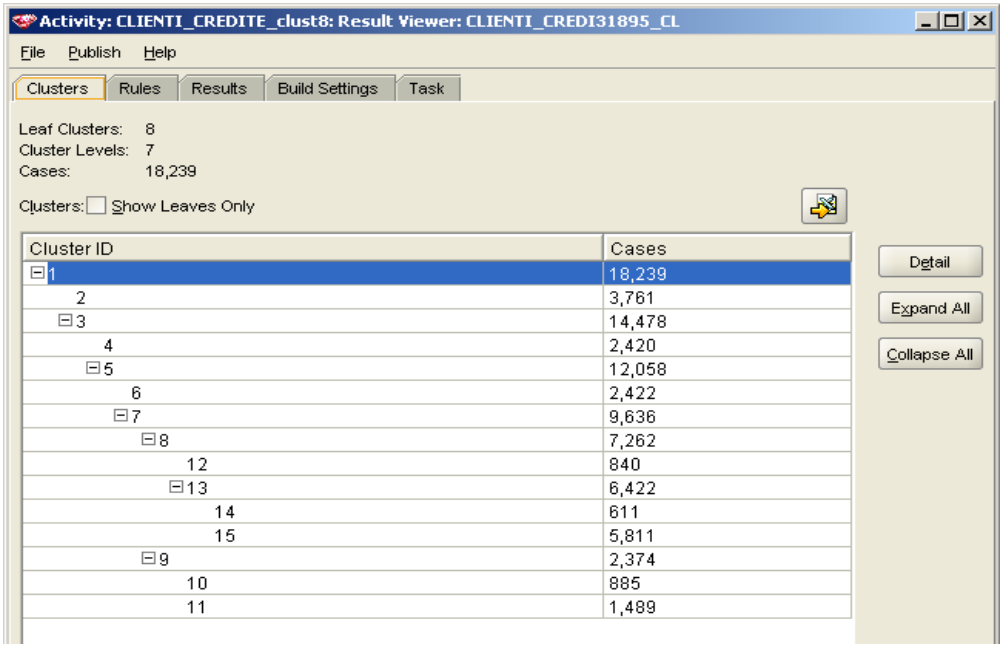


Fig. 1. Cluster distribution

Classification techniques allow us to specify by what combination of financial and demographic attributes can be characterized customers in each class and make predictions on the behaviour of new customers of the bank in order to include them in one class or another. In conclusion, in node 9 we found 2 leaves with 2 clusters, one with bad debtors and the other with good payers but with low scoring rate.

Frequency distributions of variables are represented by histograms. For example, in node 9, the amount of deposits is concentrated in the interval [600, 2142], and statistical events are focusing around 6, which means that most customers are characterized by the lowest score (ie those customers with a high probability of default). The histogram in Figure 2 shows the statistical distribution of events analyzed: 0 for the loan repayment and 1 for default. Thus, in node 9, good payers record 38% of all clients, while the debtors register the highest percentage (62%).

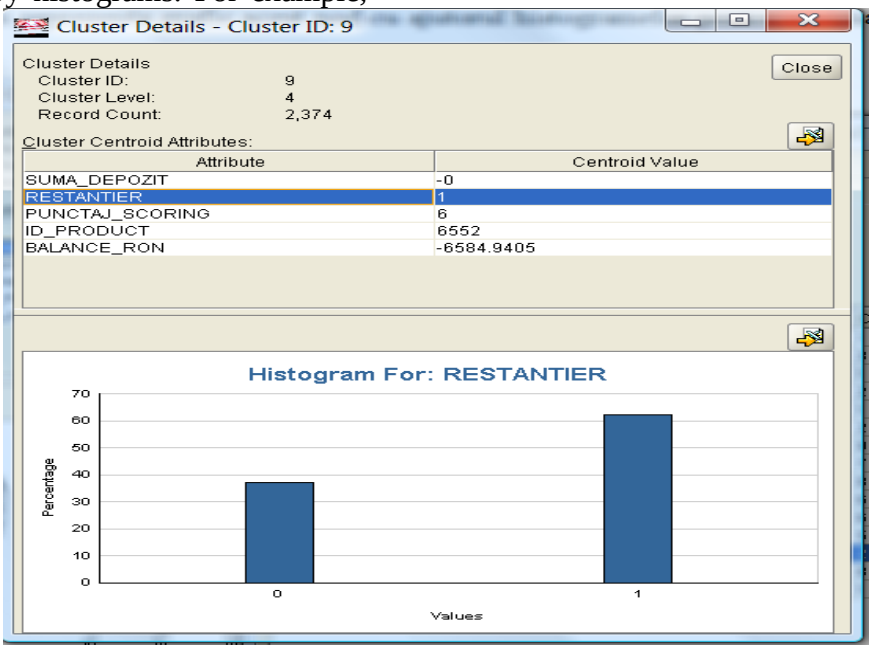


Fig. 2. The histogram for the debtor status

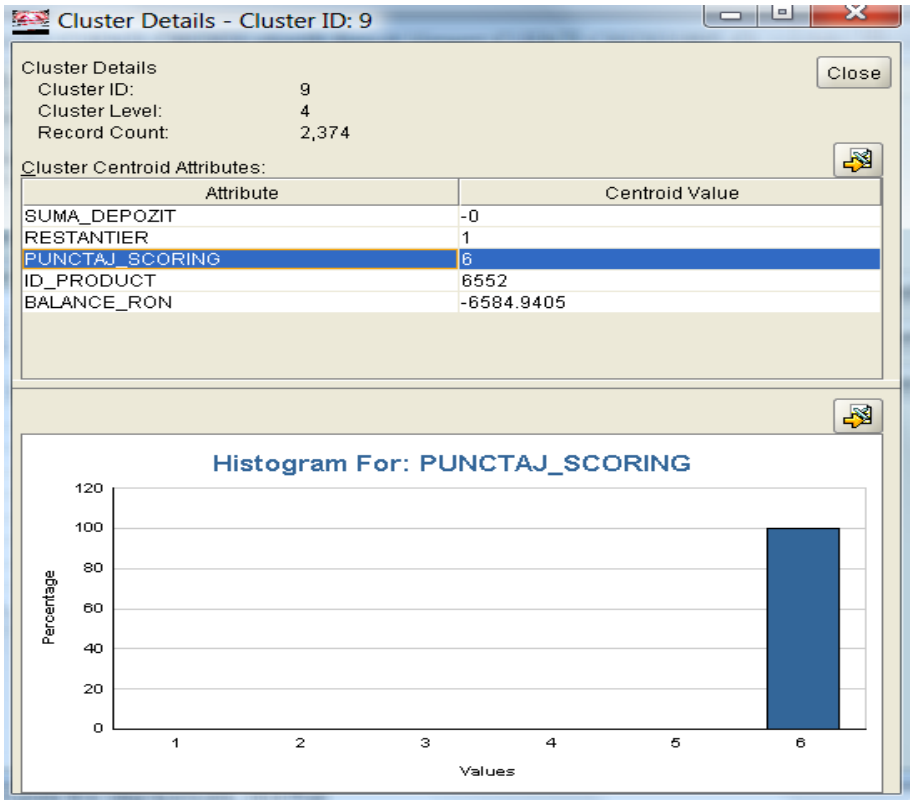


Fig. 3. The histogram for the scoring points

Figure 3 shows the distribution of the scoring rate in a node of customer data. Here, all the customers registered the weakest result when applying the scoring model. 100% received 6 points.

3.2. Applying the built model on clusters

Using cluster analysis, a customer 'type' can represent a homogeneous market segment. Targeting specific segments is an important issue of any bank's strategy. This way, each bank can develop and sell specific products and services oriented on clients needs and desires.

After applying the algorithms, we obtain the customers grouped in 8 clusters, as shown in the figure below:

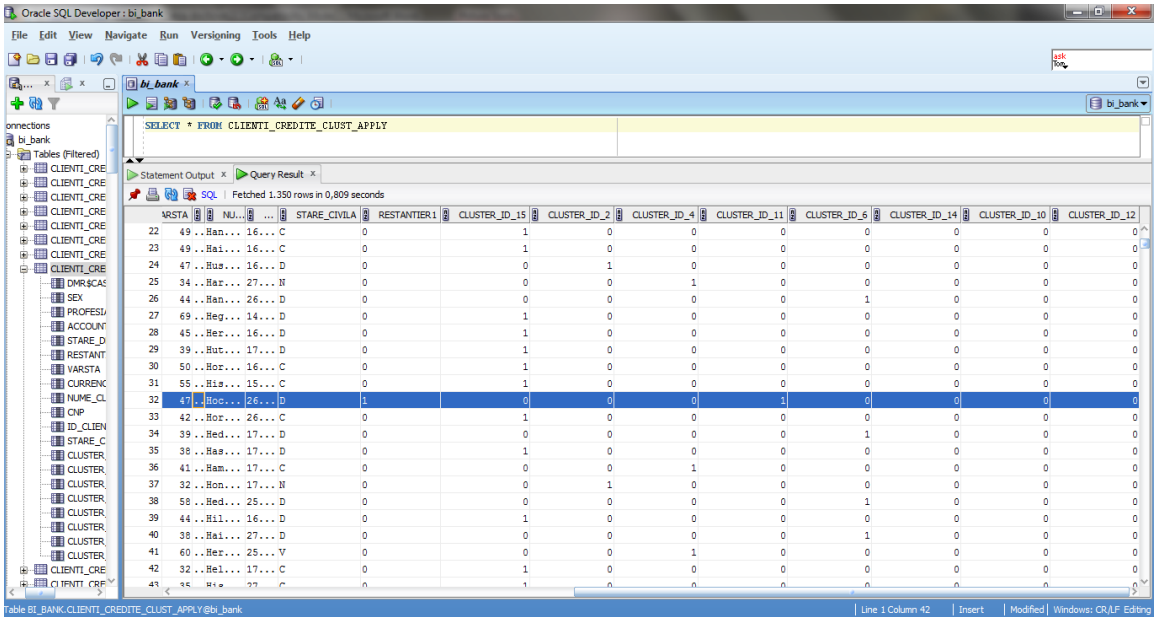


Fig. 4. Clusters situation

On these instances, we applied the predictive built models - Naive Bayes, SVM and LR, mentioning that we considered all the attributes and clusters obtained. The results are significant because all models registered over 99.8% accuracy.

Applying clustering method to our data, we identified the customers' profiles, depending on their credit history, account balance or scoring rate. Basically, the K-means algorithm was used to identify groups of customers based on scoring behaviour and divides the clients into three major classes: a) bad payers with no deposits or small amounts, who registered the lowest scoring rate (6 points) and big credit balance, b) potential insolvent clients, with scoring rate 5 or 6, small deposits and small credit balance and c) good payers with scoring rate 1-4 who have developed a good banking history.

4 Conclusions

We believe that, as of the management of massive amounts of data, data mining technology extracts successfully new knowledge from data collections, as shown in the survey presented. The built models provide decision alternatives for banking managers, based on modelling tools for the analysis of statistical data. Using statistical algorithms in order to determine patterns of behaviour, leads to creating some strong correlations for which the user is not able to generate queries. Classification and discovery of association rules are very important in the business decision process and management.

Analytical capabilities offered by the proposed solutions will produce relevant results and assist complex decision-making process, contributing significantly to improve performance in banking. Thus, an effective risk management is performed and also robust analysis, aiming to continuously improve achievement and profit margin.

Acknowledgements

This article is a result of the project POSDRU/88/1.5./S/55287 „Doctoral Programme in Economics at European Knowledge Standards (DOESEC)" . This project is co-funded by the European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies in partnership with West University of Timisoara. Also, this paper presents some results of the research project PN II, TE Program, Code 332: "Informatics Solutions for decision making support in the uncertain and unpredictable environments in order to

integrate them within a Grid network", financed within the framework of People research program.

References

[1] Han, J., Kamber, M. - *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, 2001;

[2] Hattenschwiler, P. – Decision Support Systems, University of Fribourg, Department of Informatics, DS Group;

[3] Sambamoorthi N, *Hierarchical Cluster Analysis, Some Basics and Algorithms*, 2011.

[4] Ţiţan E, Tudor A, *Conceptual and statistical issues regarding the probability of default and modelling default risk* **,** Database Systems Journal, Vol. II, No. 1/2011, pg. 13-22.

[5] Tudor A., Bara A., Botha I. - *Solutions for analyzing CRM systems - data mining algorithms*, International Journal of Computers, Issue 4, Volume 5, 2011, pg. 485-493, ISSN: 1998-4308.

[6] Ullman, J. D. - *Data Mining Lecture Notes*, 2000.

Adela Ioana TUDOR has graduated the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest Academy of Economic Studies in 2002. Two years later she graduated the MA in Financial Management and Capital Markets - DAFI, Faculty of Finance, Insurance, Banking and Stock Exchange Market. At present, she is a PhD candidate in the field of Economic Cybernetics and Statistics at the Academy of Economic Studies and also works for a leading multinational regional bank, having more than 10 years of professional experience in product management and financial analysis. Her major research interests are data mining optimization algorithms and solutions for financial institutions performance using statistical methods and techniques. During her research activity she published scientific papers and articles on OLAP technology and data mining models.

Adela BÂRA is a Lecturer at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Economic Cybernetics in 2002, holds a PhD diploma in Economics from 2007. She is the author of 7 books in the domain of economic informatics, over 40 published scientific papers and articles (among which over 20 articles are indexed in international databases, ISI proceedings, SCOPUS and 2 of them are ISI indexed). She participated (as director or as team member) in 4 research projects that have been financed from national research programs. She is a member of INFOREC professional association. From May 2009, she is the director of the Oracle Excellence Centre in the university, responsible for the implementation of the Oracle Academy Initiative program. Domains of competence: Database systems, Data warehouses, OLAP and Business Intelligence, Executive Information Systems, Decision Support Systems, Data Mining.

Elena DRAGOMIR has graduated the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest Academy of Economic Studies in 2001, pursuing postgraduate studies in Quantitative Economics and Project Management. Currently, she is PhD candidate in the field of Economic Cybernetics and Statistics at the Academy of Economic Studies and also works for a major European leasing business, having more than 10 years of professional experience in credit risk assessment, underwriting

and management of risks and enterprise risk assessments. Her major research interests are credit risk minimization models and techniques for the leasing industry, estimation models of key risk parameters, stress-testing and enterprise risk management. During her research activity she published more than 10 scientific papers and articles on methods and techniques for risk assessment and management covering both the theoretical and practical aspects of the domain.

A Framework for Semi-Automated Implementation of Multidimensional Data Models

Ilona Mariana NAGY
Faculty of Economics and Business Administration, Babes-Bolyai University
Cluj-Napoca, ROMANIA
mariana.nagy@econ.ubbcluj.ro

Data warehousing solution development represents a challenging task which requires the employment of considerable resources on behalf of enterprises and sustained commitment from the stakeholders. Costs derive mostly from the amount of time invested in the design and physical implementation of these large projects, time that we consider, may be decreased through the automation of several processes. Thus, we present a framework for semi-automated implementation of multidimensional data models and introduce an automation prototype intended to reduce the time of data structures generation in the warehousing environment. Our research is focused on the design of an automation component and the development of a corresponding prototype from technical metadata.

Keywords: *Data Warehouse; Multidimensional Data Model; Automatic Data Structure Generation; Technical Metadata;*

1 Introduction

Data warehousing projects require the employment of considerable and varied resources on behalf of the enterprises, sustained commitment from the stakeholders and a long period of development time. In most cases, these projects are successful implementations, although there are several reasons that account for failure situations, such as expenditures exceeding the budget and inability to meet delivery deadlines. Companies that undertake data warehousing development need to stay competitive on the market with regard to implementation, maintenance and other similar activities. This competitiveness is ensured significantly by small costs of development and short delivery times; therefore, reducing some of the involved costs augments the chances of successful solution delivery and increases business satisfaction.

Given these circumstances, we introduce a framework for semi-automated implementation of data structures in the data warehousing environment and propose a prototype aimed to partially automate several processes in order to reduce the related costs. Regarding the implementation of the prototype, we make use of the technical metadata provided by the underlying operational systems, external source files and the metadata generated in the data warehousing environment. We also propose a distinction between structural and logical technical metadata, and employ both these types for automating the multidimensional model (i.e. data mart) schema implementation.

Due to the complex environment of the analytical systems and the nature of multidimensional models representation, the development of such a prototype is rather challenging. The first results of our research consist in the successful initial implementation of the prototype in the SAP Business Warehouse (SAP BW) environment. In this paper we present the theoretical and technical aspects of the implementation phase, discuss the arising issues and also propose future development directions. The paper is structured as follows: section 2 presents the overall data warehousing framework architecture with its components and the interaction between them; section 3 focuses on the specifics of the multidimensional model generation

component, and also illustrates the process flow within the framework component; section 4 presents the implementation of the proposed prototype; section 5 is dedicated to the study of related work; and section 6 concludes the research and makes suggestions regarding future development.

2. Framework Presentations

The data warehouse architecture, which stands at the basis of this framework, may be decomposed into: storage elements, data handling procedures and the human factor (final users, technical staff) [1].

The proposed implementation framework addresses the semi-automated generation of the storage elements and partly data handling procedures, through corresponding ETL processes, in the data warehousing environment. The storage elements may be found in all architectural layers (data staging, data warehouse, and data mart layers). ETL processes are responsible for extracting data from source systems, transforming it in the staging area and loading the data into the storage elements for permanent storage.

Besides defining the means for automatically generating storage structures (i.e. the data warehouse and data marts schema), the framework also provides the definition metadata mapping between its components, and thus the "map" by which the automation may be achieved.

The framework provides guidance of implementing various storage structures, such as data warehouse and data marts schema, from technical metadata.

Fig. 1 depicts the proposed framework comprising the following components:

- *Data Staging Module*, which has the role of defining an integrated staging area where a single set of transformations (i.e. technical and business rules) are enforced for the entire data warehousing landscape;
- *Data Warehouse Schema Generation Module*, which enables automated generation of specific data storage metadata objects for the data warehouse layer from the technical metadata residing in the repository;
- *Data Mart Schema Generation Module*, which, similarly to the previous component, enables automated generation of the data mart schema from technical metadata;
- *Metadata Management Module*, which defines essential metadata storage and retrieval functions from the metadata repository, and handles the overall data warehousing structures generation process (e.g. replication, schema, scheduling, monitoring);
- *Data Management Module*, which facilitates the management of the data warehouse data.

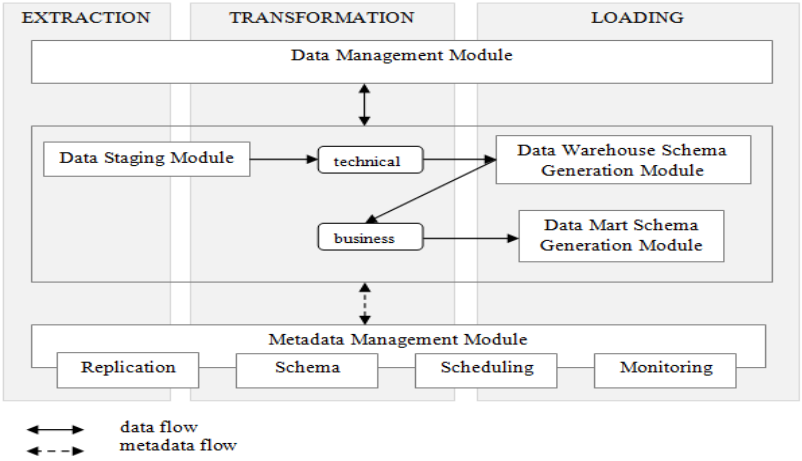


Fig. 1. Framework Architecture Overview

The components of the framework cover well-defined extraction, transformation and

loading processes in the data warehousing environment. In order to facilitate a clear understanding of the way components interact, we present a brief description of these processes in the following paragraphs.

Extraction Processes

One of the first tasks performed within ETL processes is the extraction of relevant information that has to be propagated in the data warehousing environment [2]. Extraction processes are meant to handle various and often heterogeneous enterprise-wide sources of data, which may vary from relational databases and flat files, to legacy or hierarchical model structures. The diversity of source systems foremost determines the complex nature of these processes.

Moreover, there are two types of extraction methods defined at the logical level, namely full and incremental extractions. The full extraction method assumes that all the data existent in the source systems at one moment in time is extracted completely in the data warehousing environment, whereas the incremental extraction method employs a well-defined logic by which only most recently inserted or updated records are subject to extraction. The incremental extraction (also known as partial or delta extraction) requires the implementation of advanced change capture techniques. This is usually achieved in operational source systems by logging the changes in data through a timestamp mechanism and by using change tables to track the different insertion, deletion or update operations occurred since the last extraction process. Incremental extraction is normally used on large data sets and is not applicable to external source files, such as flat files. From a physical level perspective, there are two types of data extractions: online extraction (in which data is retrieved by directly accessing the source systems) and offline extraction (in which data is staged explicitly outside the original system in flat files, archive logs or transportable tablespaces, etc. 3]).

Generally, the extracted data is temporarily stored into a physical storage environment, known as data staging area, from where data is further processed and loaded into permanent storage structures.

Transformation Processes

Transformation processes are defined as part of the overall data cleaning approaches in the data warehousing environment, among data analysis, definition of transformation workflow and mapping rules, verification, etc. [4]. Their main goal is to produce a collection of consolidated and integrated data that conforms to well-defined standards, and may provide useful information for supporting business users' decision making.

Within the ETL processes, the transformations performed are reflected at the instance level and at the schema level. The instance level transformations concern data conversions and include various individual tasks, such as cleaning of misspellings or missing data values, resolution of conflicting representations of the same data element, elimination of duplicated data records, etc. At the schema level, transformations comprise technical standardizations of data types, field lengths, etc. and semantic standardizations, such as resolution of synonyms (two or more terms refer to the same concept) and homonyms (one term refers to different concepts) 5]. Other transformation processes may include combining multiple data records into a single data record, separating one data record into multiple ones, sorting, filtering and merging operations, applying complex conversion functions or formulas, assignment of surrogate keys, etc.

Metadata, both technical and business, has an essential role in guiding these transformation processes in the data warehousing environment. Metadata is used to define data characteristics, transformation mappings, workflow definitions, etc. [4] at

the technical level, and to facilitate the understanding of converted data at the semantic level.

Loading Processes

The loading of data into the final permanent storage structures represents the last series of activities that comprise the ETL processes. Following the logical extraction methods, there are two distinct types of data loading: initial loading, normally performed the first time the data warehouse is deployed in production, and incremental loading, which represent on-going processes that occur daily, weekly, or monthly, as defined by business requirements. Once loaded into the data warehouse specific structures, the data is ready for analysis and reporting.

The presented ETL processes are part of data warehousing solution development. Their design and implementation requires extended business knowledge and correlation with business requirements. However, these processes may be to some extent automated by various generic tools available on the market, as well as by proprietary programs. In order to enable full support of various scenarios and deliver maximum value to business users, ETL tools should comply with a series of requirements, such as 6]: connectivity/adapter capabilities, data delivery and transformation, data and metadata modelling capabilities, data source and target support, data governance, design and development environment, operation and administration, architecture and integration, etc.

Having presented an overview of the overall framework architecture, as well as the various processes that stand at the basis of storage structures generation, we proceed to introducing our proposed multidimensional model schema generation component, and its implementation in the data warehousing environment.

3. Data Mart Schema Generation Module

The *Data Mart Schema Generation Module* is designed to enable faster implementation of the multidimensional data models in the data warehousing environment.

In order to achieve the desired degree of automation, we focus on capturing, storing and retrieving technical metadata within the framework's components. Moreover, as a prerequisite, we propose a differentiation of the technical metadata used and generated by the framework modules, into structural and logical metadata. Our reasoning is explained as follows:

- The *structural metadata* defines technical characteristics of the data structures, such as source systems, comprised fields, keys, data types and lengths, etc. These properties are stored in the data warehousing system independently of the way these data structures are used (e.g. as sources of data for other storage objects).
- The *logical metadata* describes the relationships (i.e. logical mappings) between the various metadata objects and storage structures of the data warehousing system, and is derived from the logical map of the frameworks components' physical model. It also comprises the logical multidimensional data model and all the other technical metadata provided by the data architect for automated generation purposes. Additionally, this type of technical metadata provides information about the data provenance and lineage (i.e. the source and path followed from source to target) [7].

When using technical metadata within the context of the defined framework, for automation purposes especially, we refer to the previously introduced two sub-types of metadata, namely structural and logical metadata.

Data Marts represent essential storage structures of departmental data in the enterprise's informational systems. They generally correspond to single business processes in the operational systems and are

built as multidimensional data models in the data warehousing environment for supporting high performance querying and optimized on-line analytical processing. Due to their particular design (i.e. star schema layout with one central fact table surrounded by several dimension tables), the number of join operations required for data retrieval is reduced to a minimum, which makes these storage structures ideal for interrogation. Moreover, data marts store aggregated and highly indexed data, supplied by the integrated and consolidated data warehouse layer. This data goes through a process of business transformations, as depicted in Fig. 1, during which more complex formulas are applied for defining key performance indicators as well as other facts of interest for the analysis process.

Among the main reasons for building data marts in the analytical environment, we mention:

- data marts are created as part of a comprehensive data warehousing solution following the development of the underlying data warehouse layer, which provides them with a foundation of highly granular, historical, integrated and consolidated data; this improves consistency and accuracy in data representation across the enterprise;
- business users or department specific requirements may be easily accommodated in the analytical environment by means of various data marts built on the solid data foundation;
- data marts lift the burden of direct analytical processing on the data warehouse data;

Considering the undeniable importance of data marts in the data warehousing environment, we propose a schema generation module that speeds the development process by automating the implementation of multidimensional data models from technical metadata. Fig. illustrates the various components that interact with the *Data Mart Schema Generation Module*.

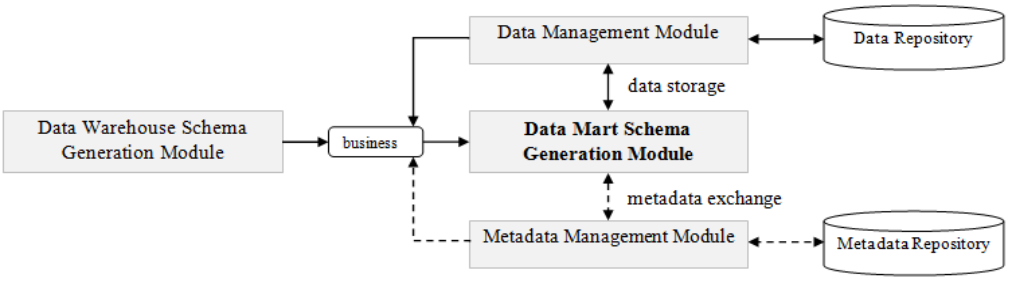


Fig. 2. Framework Architecture: Data Mart Schema Generation Module

In order to accomplish the schema generation, the proposed module exchanges technical (i.e. structural and logical) metadata with the Metadata Management Module; this consequently retrieves and stores metadata into the metadata repository. The module uses technical metadata to generate empty schemas of the multidimensional model. In the context of data mart schema generation, structural metadata is represented by the tables, columns and the data type definitions of the corresponding structures, whereas the logical metadata is represented by the multidimensional data model (i.e. fact table, dimension tables, relationships between them).

The *Data Mart Schema Generation Module* is supplied with highly granular integrated data from the data warehouse repository. However, the data is subject to comprehensive transformations based on the business logic defined by user requirements

and stored as business metadata in the repository.

Fig. 2 depicts the process flow within the Data Mart Schema Generation Module comprising six steps:

- The first step covers the loading of the XML file providing structural and logical metadata into the system. This operation is facilitated by the specifics of each implementation platform.
- In the second step, the XML file is processed and useful information is sorted out. This means that the XML file is parsed and analysed so that relevant information regarding the structure of the multidimensional data model is extracted from its content (e.g. the fact and dimension tables, the relationships between them, the fields and corresponding data types, etc.). Subsequently, the technical metadata is stored in the repository, from where it is further retrieved by the *Metadata Management Module.*
- Within the third step, the technical metadata (i.e. structural and logical) is retrieved and processed by the *Data Mart Schema Generation Module* through its interaction with the *Metadata Management Module.*
- The fourth step defines the actual schema generation process from the processed structural metadata, as indicated by the logical metadata.
- In the fifth step, the generated multidimensional schema is checked and validated against inconsistencies of technical nature specific to the implementation process (e.g. valid technical names for the fact and dimension tables, maximum number of dimensions accepted, appropriate definition of relationships based on primary-foreign key definition, etc.).
- The final step represents the communication of the schema generation results (e.g. logging of processing steps in the system, user interface, etc.).

Having defined the architectural aspects of the proposed framework, along with its main components and processes, we proceed to presenting briefly the platform - dependent prototype implementation. The technologies used for the implementation processes are: SAP R/3 as operational source systems, SAP Business Information Warehouse as analytical system, and the ABAP programming language.

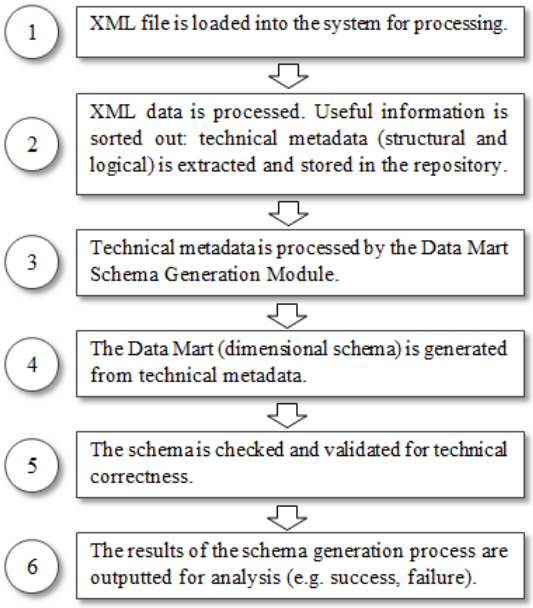


Fig. 2. Data Mart Schema Generation Process Flow

4. Framework-based Prototype Implementation

As prerequisite for prototype implementation, a Unified Modelling Language (UML) Class Diagram portraying the dimensional data model to be implemented as data mart storage structure is required. We chose the modelling of the data model schema with UML as this represents a standardized general-purpose modelling language extensively used in software engineering (aspects of dimensional data modelling with UML are presented in [8]). Moreover, numerous CASE (Computer Aided Software Engineering) tools enable the exporting of class diagrams as Extensible Markup

Language (XML) files, a commonly-used standard defined by a set of rules for encoding document.

Considering the sub-components of the automation framework, presented in Fig. 2, and their interdependence, the multidimensional data model implementation begins with the interpretation of the exported XML format class diagram from the CASE tool.

An XML parser/analyser retrieves the entities with corresponding attributes and data types, as well as the relationships between them (the multidimensional schema design), and maps the extracted information to the logical metadata represented in the repository as relational metadata tables. These represent flat relational database tables whose structure contain relevant fields, such as: entity type (fact table or dimension), automatically generated technical name, descriptive name, corresponding attributes, attributes data type and length, position in the entity, etc. The result of the XML analyser component execution is the physical storage of the multidimensional data model in relational form in the metadata repository of the analytical environment.

A subsequent action in the multidimensional generation process is represented by metadata processing, as depicted in the third step of the Fig. 2 process flow. The prototype retrieves the technical metadata from the repository (i.e. the multidimensional data model stored previously) and passes it over to the following step, namely to the multidimensional structure generation.

This data mart schema generation is implemented with by number programs, functions and system classes. Its goal is to enable the creation of platform-specific metadata objects that provide structural and physical representation of the multidimensional logical data model in the data warehouse environment.

The specifics of the implementation are listed below:

- The multidimensional model is implemented in the data mart layer of the SAP Business Warehouse, through an InfoCube structure. The InfoCube consists of several types of atomic components (InfoObjects), which form the fact table and the dimensions of the model. Internally, the cube is comprised of relational tables arranged together in a star schema, where the fact table contains key figures and the dimension tables are linked to master data tables. InfoCubes are normally supplied with data from data structures of the lower architectural levels, such as DataSources, InfoSources, master data InfoObjects, and DataStore Objects (the existence of data supply structures is ensures by the implementation of the extended version of the prototype).
- The information required for automated implementation of the multidimensional model is read from the logical metadata stored in the repository. Corresponding InfoObjects, which form the structure of the InfoCube are retrieved from the metadata repository, through a string and data type matching algorithm, as explained below (in order to facilitate this option, the data modeler should consider using the same or similar names and data types for the dimensions' attributes and a high granularity level as derived from the operational database tables).

InfoObjects represent the smallest unit of information in the SAP BW environment [9] and they are the core building blocks in the informational model (all the other data warehousing-related objects are built with InfoObjects). They are used for modelling business entities, attributes, hierarchies, key figures, or key performance indicators, as well as time, unit and currency information 10], and to create structures and tables in order to enable information modelling in a structured form within the data warehouse environment. The multidimensional generation component requires the retrieval of existing/previously generated InfoObjects

in the warehousing repository, and their use in the modelling of the InfoCube. The process is accomplished through a similarity ranking algorithm for approximating string matching [11], validated by the data type and length conditions. Once the InfoObjects have been identified and retrieved, the prototype generates the InfoCube's physical implementation.

The final step of the prototype's execution is represented by the data load and storage component's implementation. Loading processes are generated based on the logical mappings (i.e. transformations) between the InfoObjects of the data structures which represent the source of data and the target InfoCube.

5. Related Work

The automation issues in the data warehousing environment have been previously addressed in the industry and academic world, most of them referring to ETL processes design, such as data extraction and population [12], data cleaning and integration [13], [14] etc. Some research works [15], [16] propose an automation of the conceptual or logical data warehouse design. In [15] the authors propose two algorithms: one for automating DW conceptual schema derivation from OLTP schemas and another one for the evaluation of candidate conceptual schemas with user queries; in [16] the authors introduce a rule-based mechanism, which automatically generates the relational data warehouse schema by applying existing design knowledge (a conceptual schema, non-functional requirements and mappings between the conceptual schema and the source database); in [17] the an automatic generation tool for conceptual design model implementation from conceptual graphical models, which comes closest to our idea. Automation of logical and physical mappings aspects in the data warehousing environment, as treated in [18], [19], [20] have also been considered in our approach.

However, to the best of our knowledge, automating the creation of data warehouse-relevant data structures from technical metadata, with the aim of defining a comprehensive enterprise wide solution has not yet been covered. We argue that given the existence of external metadata files provided by data modellers, our proposed prototype shortens considerably the manual repetitive work of developers and enables the implementation of an initial enterprise data warehouse model.

5 Conclusion

The framework architecture for semi-automated implementation of multidimensional models and the corresponding prototype, presented in this paper are designed for providing faster development of data warehousing projects, and reducing related costs and delivery times. Our main contributions consist in the design of a framework that defines several automation components and the boundaries of the systems for its implementation. The main goal achieved by the corresponding prototype implementation is the interpretation of UML class diagrams exported in XML format and the semi-automated generation of the multidimensional model in the data warehousing environment. The proposed prototype implements the physical data model from existing and generated technical metadata and logical mappings of the underlying metadata repository.

Several enhancements are to be made in the development of the prototype; therefore, we continue our research in order to improve the functionalities, enable integration of additional data sources types, and optimize the execution processes.

6. Acknowledgment

Ph.D. scholarship, Project co-financed by the SECTORAL OPERATIONAL PROGRAM FOR HUMAN RESOURCES DEVELOPMENT 2007 – 2013;

Priority Axis 1. "Education and training in support for growth and development of a knowledge based society"; Key area of intervention 1.5: Doctoral and post-doctoral programs in support of research. Contract nr: POSDRU/88/1.5/S/60185 – "Innovative doctoral studies in a Knowledge Based Society" Babeş-Bolyai University, Cluj-Napoca, Romania.

References

[1] J.A. Rodero, J.A. Toval, and M.G. Piattini, "The audit of the Data Warehouse Framework*," in *Proceedings of the International Workshop on Design and Management of Data Warehouses*, Heidelberg, Germany, 1999, pp. 14-1:14-12.

[2] D. Theodoratos, S. Ligoudistianos, and T. Sellis, "View selection for designing the global data warehouse," *Data Knowledge Engineering*, vol. 39, no. 3, pp. 219 - 240, December 2001.

[3] P. Lane, *Oracle Database Data Warehousing Guide, 10g Release 2 (10.2)*. Redwood City, CA: Oracle, Inc., 2005.

[4] E. Rahm and H.H. Do, "Data Cleaning: Problems and Current Approaches," *IEEE Data Engineering Bulletin*, vol. 23, no. 4, pp. 3 - 13, 2000, http://sites.computer.org/debull/A00DEC-CD.pdf.

[5] P. Ponniah, *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*. New-York: John Wiley & Sons, Inc., 2001.

[6] T. Friedman, M. A. Beyer, and E. Thoo, "Magic Quadrant for Data Integration Tools," Research Report ID Number: G00207435, 2010.

[7] M.H. Bracket, *The Data Warehouse Challenge: Taming Data Chaos*, 1st ed.: John Wiley & Sons, 1996.

[8] M. Muntean, "Implementation of the Multidimensional Modeling Concepts into Object-Relational Databases," *Revista Informatica Economică*, vol. 3, no. 43, 2007.

[9] SAP AG. (2009) Sap Library Document. [Online]. HYPERLINK "http://help.sap.com/SAPhelp_nw04/helpdata/en/" http://help.sap.com/SAPhelp_nw04/helpdata/en/

[10] K. McDonald, W.H. Inmon, A. Wilmsmeier, and D.C. Dixon, *Mastering the SAP Business Information Warehouse*. Indianapolis, Indiana: Wiley Publishing, Inc., 2002.

[11] S. White. (2004, February) Catalysoft. [Online]. HYPERLINK "http://www.catalysoft.com/articles/StrikeAMatch.html" http://www.catalysoft.com/articles/StrikeAMatch.html

[12] J. Adzic, V. Fiore, and S. Spelta, "Data Warehouse Population Platform," in *Proceedings of the VLDB 2001 International Workshop on Databases in Telecommunications II* , London, UK, 2001.

[13] M. Jarke et al., "Improving OLTP data quality using data warehouse mechanisms," in *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, Philadelphia, PA, 1999.

[14] V. Tziovara, P. Vassiliadis, and A. Simitsis, "Deciding the Physical Implementation of ETL Workflows," in *Proceedings of the ACM tenth international workshop on Data warehousing and OLAP*, New York, USA, 2007.

[15] C. Phipps and K. Davis, "Automating Data Warehouse," in *Proceedings of the 4th International Workshop on Design and Management of Data Warehouses*, London, UK, 2002, pp. 23-32.

[16] V. Peralta, A. Illarze, and R. Ruggia,

"Towards the Automation of Data Warehouse Logical Design: a Rule-Based Approach," in *Proceedings of CAiSE Short Paper Proceedings*, 2003.

[17] K. Hahn, C. Sapia, and M. Blaschka, "Automatically Generating OLAP Schemata from Conceptual Graphical Models," in *Proceedings of the ACM DOLAP 2000*, 2000, pp. 9-16.

[18] S. Benkley, J. Fandozzi, E. Housman, and G. Woodhouse, "Data element tool-based analysis (DELTA)," MITRE Technical Report 1995.

[19] M. Castellanos, A. Simitsis, K. Wilkinson, and U. Dayal, "Automating the loading of business process data warehouses," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* , 2009, pp. 612-623.

[20] E. Rahm and P.A. Bernstein, "A survey of approaches to automatic schema matching," *The VLDB Journal — The International Journal on Very Large Data Bases*, vol. 10, no. 4, 2001.

Ilona-Mariana NAGY is currently a PhD Candidate in Cybernetics and Statistics at the Faculty of Economics and Business Administration of the Babes-Bolyai University. She has a

Bachelor's degree in Business Informatics and a Master's degree in Business Informatics and the Informational Society. Her main scientific fields of interest include: databases, Data Warehousing, Business Intelligence and SAP technologies.

Analysis on Cloud Computing Database in Cloud Environment – Concept and Adoption Paradigm

Elena-Geanina ULARU [1], Florina PUICAN [2], Manole VELICANU [3]
[1, 2] Phd. Institute of Doctoral Studies
[3] Professor, Academy of Economic Studies Bucharest
ularugeanina@yahoo.com, puicanflorina@yahoo.com, manole.velicanu@ie.ase.ro

With the development of the Internet's new technical functionalities, new concepts have started to take shape. These concepts have an important role especially in the development of corporate IT. Such a concept is „the Cloud". Various marketing campaigns have started to focus on the Cloud and began to promote it in different but confusing ways. This campaigns do little, to explain what cloud computing is and why it is becoming increasingly necessary.
The lack of understanding in this new technology generates a lack of knowledge in business cloud adoption regarding database and also application. Only by focusing on the business processes and objectives an enterprise can achieve the full benefits of the cloud and mitigate the potential risks. In this article we create our own complete definition of the cloud and we analyze the essential aspects of cloud adoption for a banking financial reporting application.

Keywords: *Database, Cloud Computing, SaaS, IaaS, PaaS, Virtualization*

1 Introduction

Probably the most revolutionary technology of the last decade developed in the Internet is represented by the Cloud. This technology has started to develop due to the promises of developing a futurist business environment. In this kind of environment each company is supposed to spend the minimal amount of money on resources and gain a lot from the work developed. The promise of Cloud Technology consists also in minimizing the Total Cost of Ownership (TCO) making it right for leaders and CEOs to dream about a modern and futuristic company. The idea that a "successful businesses may soon have no chief executive, no headquarters and no IT infrastructure" (Dr. James Bellini) has started to appear in each leader's mind. These are the directions by which Cloud Computing is leading the future. [3]

In order to be able to understand the necessity of this technology we have to define it first. The cloud concept is spoken, written and used in many other ways very often by different people belonging to diverse fields especially economic and technological. We can realize that by reading a variety of IT blogs, hearing discussions or taking part at conferences. However, the definitions differ from one another. Why there are so many definitions that support misunderstanding about cloud computing?

In our opinion all the perspectives on Cloud computing depend on the background of each specialist. For Dr. James Bellini as futurologist cloud computing is projecting the image of a company with no chief executive, no headquarters, no database on their own machines and no IT infrastructure. "The most valuable resource of this new king of company it will be represented by its connected eco-system" he says.

The idea that Dr. James Bellini encompasses in his statement defines cloud computing as an evolving paradigm. The idea is also shared by economists who complete it with the economical features. "The Economic Benefit of Cloud Computing" publication defines cloud computing as follows: *"Cloud computing uses the same paradigm of shared needs, costs and shared resources leading to shared savings since duplicate ancillary costs (e.g., facilities, power, a/c, personnel, etc.) are greatly reduced, if not*

completely eliminated. Add to these obvious savings areas the virtual elimination of the capital expense associated with annual software licensing (e.g., initial license purchase, annual maintenance, individual help desk support, etc.) and the user savings become dramatic and easily justified. This is especially true in this era of increasing IT needs coupled with decreasing IT budgets. Cloud computing represents one of the few means of meeting next year's IT requirements with last year's budgetary amounts."[7]

This above definition justifies the economical importance of Cloud Computing and states the major advance in IT and business represented by this technology. Speaking from the global point of view it is based on the major resource provider's economies of scale, such as Google. The quality of services is accompanied by massive cost reductions, often over 50%. Basically a major economic advantage of cloud computing is the fact that "it turns capital expenses into operating expenses (CAPEX to OPEX)". Through this reversing of expense type companies are able to direct their capital to other business investments different from IT.

The Cloud is a solution that provides new ways of using external resources that you can configure according to your needs - servers, storage, applications and services. A great economical advantage is that resources are leased by the Cloud provider on-demand and in variable quantities so that the client only pays what he consumes. In other words, if you do not use the resources, you do not pay it.

The secret is the emerging technology that allows all types of companies to match the technical resources of multinationals. To implement a system within a company needs a lot of resources both material and human. If we're referring to a small business, these investments are limited and any IT investment affects the investments needed in other business areas. That is why Cloud is an ideal solution for small or medium businesses, providing companies the resources to hire an insignificant cost, so that small companies can match the technology of one of the most advanced technologies in the world, which is available locally. The real advantage is that anyone can use for free or for a small amount of money the best services on the market. Companies of 3 or 3,000 people can use immediately cloud services and the price will be directly proportional to the use of the resource. Cloud solutions spare the costs for servers, software licenses, hosting, collocation, database maintenance, specialized technical personnel, upgrades of all kinds, annual subscriptions, etc.

You just need the final service and you can benefit from the Cloud. We assume that we all have a computer with access to the Internet, so we can instantly be Cloud services users. For example, Gmail now allows storage of 25 GB of emails. A local solution in-house offering lower performance would cost at least 3-4 times more per year.

Another definition, written from an academic perspective is presented in the article "Above the Clouds: A Berkeley View of Cloud Computing" wrote by the researchers of RAD Lab (Reliable Adaptive Distribute System Laboratory) at the University of Berkley. This article points out the most important features of cloud computing: *"Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we will call a Cloud."* [1]

The definition explains the essence of a Cloud and the considered component elements. It refers both to the software and hardware, introducing a new term "Software as a Service" that will be further explained.

In our opinion the above definition is an attempt to create a common perspective on Cloud Computing for all the interested parts. Standardization is very important. The lack of standards in information technology can make a technology trickier to use that is why standardization is a goal in every attempt to define the Cloud.

An effort of standardization for the cloud computing definition of made by The National Institute of Standards and Technology (USA): *"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models."* [2].

Despite the fact, that cloud computing is still evolving, NIST definition is accurate regarding the solutions nowadays provided. Cloud is not only the next step in computing development, it is a new perception on business applications and the addressed processes, consisting in the movement of all data and computing power on the Internet.

The Complexity of enterprise systems increases with each new technology implemented. As changes in the business environment occur so appeared the need to implement more efficient enterprise systems consisting in great investments of money and human resources. In most cases within a company there are a lot of enterprise applications developed in multiple technologies and also the related databases that the company is forced to have in operation because of investments made over time.

Because these applications justify their existence only by completing the tasks for which they were implemented without outstanding performance, every company wants to invest in the latest technology so it can enjoy the economic benefits of the digital economy and increased performance gained through innovation. This is where the Cloud intervenes. Through Cloud adoption, each company can benefit from the functionality of the former application and in the same time the performance of the Cloud.

Considering the above explanations about what cloud computing really is, here comes the definition of this article – taking into consideration the standards, and our practical and theoretical knowledge: *"Cloud Computing represents a technology concerning the application as service and a shared pool of configurable computing resources over the Internet based on on-demand system of providing with minimal effort and provider interaction. It is a way to cut the costs, speed-up deployment of new technology and move responsibilities from managers to the cloud provider"*

This definition contains some of the most important characteristics of cloud computing, and some main features that come with the implementation and deployment of cloud services. It combines different views in forming one in order to be the most simple and useful so it can be used as a definition that most people understand, not only IT professionals.

In order to be able to fully understand the cloud we should forget what we've been taught before about traditional IT when the servers where residing in the same location with the company because the inflexibility of conventional IT is surpassed by the cloud computing service model.

2. Cloud Characteristics

The term "Cloud Computing" most likely derives from the cloud's diagram used to represent the Internet. The concept started when telecommunications companies have made a radical shift from point-to-point connections to VPNs (Virtual Private Network) in 1990. Optimization of the use of resources through more efficient load balancing work has brought them major

savings. The term, Cloud Computing was first used in 1997 at Ramnath Chellappa's lecture, when he defined it as "a computational paradigm where computational limits will be determined by economic reasons and not technical limitations".

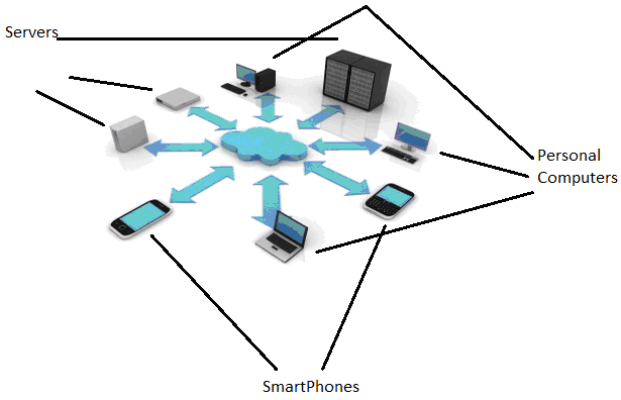


Fig. 1. Proposal for the Graphical Representation of a Cloud

The essence of the Cloud Computing concept is mainly in terms of hardware, with deep consideration to software also, because the processing power is moved on the Internet. Thus some servers placed at addresses unknown by the user bring processing power without physical possession, but only virtually by renting their computation and storage capacity.

The Cloud Computing evolution in the latest years is one of the greatest advances in computing history. Given that this technology will reach its potential in several years, we must have a sound understanding of technology both in terms of consumer and provider.

Cloud Computing is making a lot of advances into all aspects of IT. Databases in the Cloud will experience major architectural changes to take advantage of massive scalability and large amount of data. A product which was released by Oracle especially for database control on the Cloud is Oracle Enterprise Manager 12*c*. This tool is known for "creating business value from IT by leveraging the built-in management capabilities of the Oracle stack for traditional and cloud environments, allowing customers to achieve unprecedented efficiency gains while dramatically increasing service levels."[8]

Although many studies are now based on the technology itself, there is a great need for research of the business implications for cloud adoption. In this paper we describe the key issues for understanding the concept of cloud computing.

Cloud service models

The Cloud Computing concept is strictly defined in this article through the services of three models. The **service model** describes an approach whereby the cloud supplier aim to satisfy clients' demands.

Cloud Computing provides functionality to users as a service. This technology changes the angle that we were looking upon software and hardware. The owner of a business should not invest in physical devices. Thus, he must use cloud resources as an alternative solution through pay-as-you-go payment method so he can pay only the resources that he needs. The system recognizes subsequent requests and provides customer the performance they require. Although all cloud services work in the same way they differ in the abstraction level, thus being classified in the following 3 types of services: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). [6]

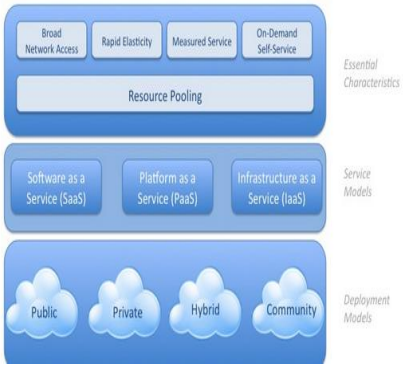


Fig.2. Visual Model of the NIST definition provided on the concept of Cloud Computing [2]

Software as a Service (PaaS)

Software as a Service (SaaS) is a software service model that provides client remote accesses to business functionality such as services offered by cloud. The customer does not purchase software licenses. The infrastructure cost, the maintenance, the rights to use software support are all gathered in a monthly payment or pay as use service. The application can be accessed via the Internet, the most common web browsers that do not require any special hardware or software drivers. It can also be independent of the computer operating system of the machine that uses the cloud services. Saas is very similar to web applications, spread across three levels: user interface, business logic, and data. On the other hand SaaS applications have special functionalities involving billing and measurement methods based on pay-as-you-go type of payment.

A major advantage of using SaaS applications compared with local machines is the fact that the Cloud applications are not integrated into business applications. Firstly, when you want to change supplier, it's all about how quickly can managers find a SaaS product with more capabilities. There is no need to redesign the application, analyze it in-depth or worry about application maintenance. What you need to do is register on the new provider's website, provide access to employees, set the information for payment and the application is ready to use on the Cloud.

A confusion that is often seen is the misunderstanding created between the concept of SaaS and SOA (Service Oriented Architecture). SOA is a methodology and a framework that defines how services should be delivered to serve certain functionalities. To differentiate the two concepts we can compare SOA with a strategy and SaaS with one of the tactics used to reach targets in the strategic plan. SOA is a term more abstract and cannot be replaced with SaaS or vice versa.

Saas is a way of delivering applications. The concept of having applications built up in a database supposes to have the freedom to access, without installing any complex software or hardware. The virtual databases maximize the computing resources and it improves the ability to predict resource usage. We suggest that the operations of the database within Saas to be written in such a way to maximize concurrency and minimize exclusive locking. Databases could be in terms of shared schema for Ad-hoc, Custom Configuration maturity models - shared databases, but also separate databases for scalable, configurable, multi-tenant maturity models.

Platform as a Service (PaaS)

PaaS is a cloud service model working at a lower level of abstraction, compared to SaaS. The functionalities offered to customer are to develop and maintain start-up cloud based applications. The client has no access to infrastructure features such as network configuration, CPU performance, storage or operating systems. Infrastructure is strictly determined and is not in the

client's control. There are still some settings that can be changed and customized according to customer preferences [2]

PaaS is usually described by three groups of services: basic services, infrastructure services, application services. The basic services contain items such as operating systems, storage systems, file and database. All these elements are to be customized by the client. Infrastructure services include authorization, authentication, security settings and settings for online storage. The third group of services, application services refers to functionalities such as measurement and billing. Cloud platforms are very similar to business platforms except that they are made to operate on the Internet.

Although the PaaS model is not as famous and implemented as SaaS, according to Forbes, the year 2012 will be the year that PaaS will be used to its maximum potential. [5].

Our proposition of databases within PaaS concept imply to have open source databases with high performance in replication than native solutions for MySQL and also continuous operations in order to enable zero-down database upgrades. In such a manner the database performance scalability solutions will offer database high availability and incremental performance scaling using commodity databases.

Infrastructure as a Service (IaaS)

IaaS is a cloud service model working at the lowest level of abstraction. Its main features are the provision of processing, storage space, network and other fundamental resources. The hardware infrastructure is outside the control of the client. However he is able to run different software. Sometimes the customer has also the option to change some hardware settings so as to optimize the solution. [2].

Many IaaS providers utilize shared databases. In the case of any application needed to be virtualized and run in an IaaS environment there is a need of a relational database. Our proposal implies to have such databases in IaaS, as they provide the best solutions experimented until now.

3. Cloud Database Adoption

We already know that Cloud Computing is here to improve the capabilities of a business but how do we know which is the best way to adopt cloud or if the activities in our company are ready to adopt a cloud solution. Taking into account different aspects of our business activity we have to make a thorough assessment before taking the cloud adoption decision. From out work experience we are going to analyze and asses the cloud adoption decision in a banking environment. Banking applications are highly based on databases and the quality of the database administration reflects onto the application itself. Whether you are assembling, managing or developing on a cloud computing platform, you need a cloud compatible database. This paper presents a use-case that shows business problems addressed by using Cloud computing, describing the business considerations that influence an organization to use Cloud computing. It is built on the structured used by The Open Group and enhanced with our own considerations.

This business use-case template is comprised of six main elements:

1. Category – industry, sector, or environment – the category is very important due to the particularities that each of this category has
2. Company Background
3. Business Problem/Description of the situation – the migration to cloud can be suitable only for certain business
4. Actors impacted by this business problem –is very important to know which are the impacted parties when adopting the cloud
5. Key Business Requirements (encompasses quality and/or service-level requirements – where known)

6. Business Risks experienced by this problem – referring to database security, data loss etc, risks [4]

The following table contains a use case for migrating a financial reporting application database in the cloud. The industry/category referred here is represented by the banking environment and specifically by an Eastern European bank with a local representative in Romania. The use case consists in migrating the bank's financial reporting application and database in the cloud.

Category: Banking - Financial Reporting

Company Activity Background

- The bank here in discussion is a local bank forming part of an international group. It needs to report last month's financial situation and provide this information to the mother bank. The bank must address the need to produce different reports with monthly and yearly financial situations to business directors from within the bank. It must also hold a historical database with all the previous reports.

Business Problem/Description

- This company lacks the necessary in-house skills and resources to support the high degree of sensitive data that is required to support its key business processes. Also the financial system already implemented for the financial reporting is growing on data and needs more storage and processing power. The data needs to be aggregated, filtered, and assembled in order to be more useful and generate value for the target audience. Database administration is a problem because of the lack of specialists in the bank.

Actors

- Bank Directorate
- Operational Management
- Business –Controlling- owner of the application
- IT Operations Management
- Database Administrators

Business Requirements

- Need to generate financial reports for directors and to provide on time banking financial situations reporting to the mother bank. Also from the IT Operations Management -ensure that business activities can be carried out during times of additional complexity (reporting period, running of the application such as calculating transfer pricing for whole month data). The database needs to hold accurate data for reporting purposes.

Business Risks

- We have divided this risks in two because they connect and are related: Database risks: Current business operation is exposed to disaster and recovery backup processes either being insufficient or not in place; Database errors are frequent and hard to solve.
- Business risk : Lost information contributes to poor business decisions.
- Loss of business continuity across processes and new ventures; Significant loss of assets or revenue flow; Loss of transparency for stakeholders Inadequate governance due to lack of timely and inefficient reporting.

Modernization

Company Background

- This bank must address modernization of its reporting's processes and applications. However, they do not have the skills in-house to conduct such activities. Existing business services and assets are out-of-date and need modernization. They are using an old database processing system.

Business Problem/Description

- Business needs to have performance in the reporting process and to shorten the reporting time from 15 banking days to less than 10.Business does not have the skills in-house to transition via Cloud facilitation to modernize business processes.

Actors

- Business Procurement Management

- Controlling Department
- Project Management
- Data Center Management
- IT Operations Management
- Database Administrators

Business Requirements

- Need to modernize infrastructure
- Need to modernize applications
- Need to modernize business processes
- Improve employee skills
- Better performance

Business Risks

- Lose of performance
- Losing opportunities due to inability to support new technology
- Impact on existing contracts for support, licensing, and services. Limited licenses for database related products.

Rapid Business Capacity & Scale

- Company Background
- This business needs to scale up its operations rapidly, including increased IT capacity – within a short cycle of days to a few months – to meet specific operating workloads.In particular the bank needs to develop a daily reporting system similar to the monthly one. Developing this system will need a tasting environment similar to the production one. The capacity needed will be covered by a lot of purchases translated into EXPENSE.

Business Problem/Description

- There is deficiency of resources and capacity to meet business activity demand to support day and night time peak loads. The company is facing issues such as how to balance compute workloads better; how to optimize costs of operations; and how to follow variable demands of service effectively and efficiently at lower cost.

4. Conclusions

As conclusion, we strongly recommend enterprises to adapt cloud services no matter the field they're in. Before that however it is necessary to follow and analyze every issue according to particular business and finally deploy cloud services based on that analysis. In cloud adoption as we stresses above it is very important to know the essential elements of the technology that you want to have in your company and on these core elements analyze and build your own business cloud functionalities.

References

[1] M. Armbrust, A Fox, R Griffith, A. D. Joseph, R. H. Katz "About the Clouds: A Berkeley View of Cloud Computing", *Electrical Engineering and Computer Sciences. University of California at Berkeley, February 10, 2009*, pp. 2-5.

[2] P. Mell, T. Grance, The NIST Definition of Cloud Computing (Draft), National Institute of Standards and Technology, January 2011.

[3] The Telegraph, *"In the future, Cloud Computing will be the only choice*", August 2011.

[4] P K. Isom, D Hawley, *"Strengthening your Business Case for Using Cloud",* Strengthening your Business Case for Using Cloud, July 2010.

[5] T. Nielsen, "2012: *Time To Adopt Platform As A Service*", February 2012.

[6] B. Sosinsky, "The Cloud Computing Bible", Wiley Publishing, 2011.

[7] K. L. Jackson, R. Williams, **"***The Economic Benefit of Cloud",* Executive White Paper, Clear Government Solutions, September 2011

[8] Oracle Corporation site, „ Overview of Oracle Enterprise Manager Cloud Control 12c" [Online] Available at: http://docs.oracle.com/cd/E24628_01/doc.121/e25353/overview.htm.

Geanina Ularu graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2009. She is currently a Phd. student at the Institute of Doctoral Studies from the Academy of Economic Studies Bucharest. Her thesis title is "Optimization solutions in the Cloud".

Security Aspects for Business Solution Development on Portal Technology

Ovidiu RĂDUTĂ, Adrian MUNTEANU
Institute for Doctoral Studies within Academy of Economic Studies
Bucharest, ROMANIA
ovidiu.raduta@gmail.com, adrianm21@yahoo.com

In the scope of portal development, in order to talk about security issues, concerns, and solutions, it is necessary to define a few terms: authentication, authorization, Single Sign-On (SSO), confidentiality, integrity, and non-repudiation. Focusing on the scope of what the portal developer and designer need to know, below it will be explained these concepts, considering it is important to define and make a brief analysis of these terms for understanding of achieving the security goals.

Keywords: *Role-Based Access Control, Data Encryption, Data Integrity, Non-repudiation, Grid Computing*

1 Introduction

In recent years, more and more companies have been creating portals in order to offer to their employees, partners and customers to access critical and sensitive information. Most portal developers have applied high level security challenges in this regard. Many vendors have created particular interrelated solutions for some of the security requirements, trying to tie developers to a particular product. Generally, this is a good choice and often works well until the product doesn't work [11]. In this case, the software is no longer supported, or you need to readjust the architecture. Of course that a developer who has been in that boat feels his pain because he has been there. It appears a question: How can portal developers meet their security challenges with a standards-based, open-source security solution? This paper answers that question – first, presenting some essential security concepts, and then by highlighting you standards, techniques and open-source tools that it can use in order to secure your portal solution [8].

Second, this paper will talk about Grid portals, which are an increasingly popular mechanism for creating customizable, Web-based interfaces to Grid services and resources. Due to the powerful, general-purpose nature of Grid technology, the security of any portal or entry point to such resources cannot be taken lightly, in particular if the portal is running inside of the trusted perimeter, such as a Science Gateway running on an SDSC machine for access to the TeraGrid.

2. Core Security Concepts.

First, we will talk about the main concepts regarding portal security

2.1. Authentication

Authentication is the first point in providing access control and this involves validating the identity of a user. Most of cases in a portal environment, authentication may be achieved through user name/password login, validating a user's client certificate, or through validation via smart card or biometric device.

In order to proceed, authentication it is necessary to develop a solution which is usually based on a repository for validating these identities and integrating it with the system. Regarding it, a mutual authentication means to provide the identification for both of parties involved in communication, and this is done using a particular security protocols, such as SSL/TLS. To ensure that the message was sent by the expected sender, it is used a

message origin authentication which is not replayed. [15]
Being one of the most important aspects to providing security, authentication at the portal level will dictate how your application interacts with other enterprise applications and Web services. [7]

2.2.Authorization

When an user is validated, the next step we have to assume is what the user has permission to do. The access control separation in two distinct mechanisms, authentication and authorization, provides a logical separation of first validating identity, and then validating what resources that entity has access to consume or produce. Authorization defines the user permissions, roles, and other useful credentials which are used to permit access to certain portal services. An access control strategy – Role-Based Access Control (RBAC) is provided – (useful because of capability is prevalently in J2EE architectures).
In extension, RBAC (as framework for the authorization management of credentials) is essentially useful in many portals, relational databases, or commercial web-based systems. As it can be seen in Figure 1, a key component of RBAC maps roles to permissions, and maps users to roles. There are also shown views of some authorization mechanisms of the traditional access control – using Access Control Lists (ACLs). So far, large and complex ACLs made links users-permissions, and restricted access to resources by permissions and users. Because there are many cases when the number of permissions is usually very high, an access control lists for discretionary access can be difficult to be kept (with subject -object mappings). So, generally, abstracting the user and the resources permissions is very useful in authorization management. [15]

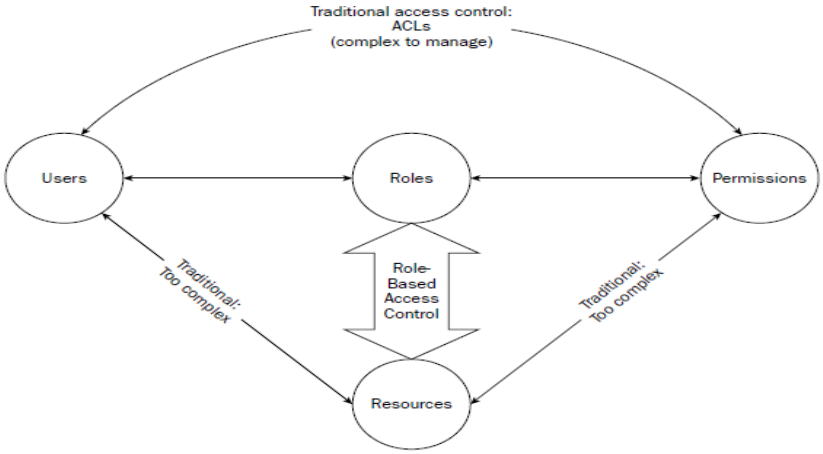


Fig. 1. Key component of RBAC

Although, in a certain organization, mapping these permissions to never-changing roles can happen at one time and users can be assigned and unassigned to these roles during the lifetime of the individual, making the access management control easier to keep. [9]
I consider that the most difficult part of setting up a role-based access control is that an organization must define its roles based on the proper processes. They consider that the technical part of the solution is the easiest phase. They have to provide an enough flexible schema in order to be able to cope the reorganization's rigors. With "Introduction to Role-Based Access Control", NIST (see references) provides an appropriate explanation: "*With role-based access control, access decisions are based on the roles that individual users have as part of an organization. Users take on assigned roles (such as doctor, nurse, teller, manager). The process*
of defining roles should be based on a thorough analysis of how an organization operates and should include input from a

wide spectrum of users in an organization." ([10], pp1)

2.3.Confidentiality

Keeping the information secret it is mandatory when sensitive information is sent. Confidentiality is considered the security goal for hiding information and encryption it is an appropriate solution to provide it. With encryption, a plaintext message is modified with a cryptographic algorithm in order to produce a ciphertext message. In the same time, it is a must to have the possibility to decrypt the data using a key. Many different encryption/decryption algorithms, even if it is about a symmetric algorithms (secret-key) or an asymmetric algorithms (public key), can be used to offer different protection levels of sensitive data. Different things like key management for distributing keys, ciphers to use, and cryptographic protocols that provide these services, are needed to create solutions in order to satisfy confidentiality requirements. [12]

Many higher-level protocols like Transport Layer Security (TLS) and Secure Sockets Layer (SSL) (its later version), offer bulk encryption between two points. At this level, the cipher is determined and the encryption key is known at the beginning of the protocol in order to establish a "shared secret" understood on both sides. It is good to know that SSL is a point-to-point protocol which can be used for one-way or two-ways authentication between two points only. Generally, it is enough such a session for environments with a simple client and server in order to protect the data confidentiality in the transmission.

Encryption requirement. If you have some encryption requirements, the portal should be able to put it up, but the portal architect has to foresee that the cryptographic mechanisms will impact the performance. [13]

Requirements. Generally, to have requirements satisfied, it is enough to have a simple SSL/TLS connection between the user and the portal. If not, that means it is necessary to be size encryption between the user, the portal, and all nodes, web services involved and enterprise applications in the solution. (thereby, the solution becoming a bit more difficult).

Secret data. If it is needed to keep some confidential information away from web services (the encryption needs to be directly between the portal and the eventual data source), you will need a shared secret between the portal and the data source (not just bulk encryption between the nodes). Here, XML Encryption, a W3C standard, being quite useful for this. [12]

2.4.Data Integrity

In transit, ensuring a non-altered data is a must. To validate the integrity of a message means to use techniques in order to prove that data integrity is kept. Because on a TCP/IP network could occur some message injection, packet tampering, or IP spoofing, many applications require a digital signatures, a MAC (Message Authentication Codes), or hash algorithms.

The portal's architecture offers integrity challenges regarding users and the portal, and also between portal, web services and the enterprise data source. Beginning from requirement set, SSL/TLS may provide it (message integrity between users and portal). In addition, other standards can be used to achieve integrity. One example is XML Signature (a W3C standard), which provides message integrity in addition to non-repudiation (see the next paragraph). A mechanism that achieves such a solution for Web services is provided by OASIS Web Services Security specification.

2.5. Non-repudiation

The side effect of digitally signing a message is called non-repudiation, which is a security service used when a user has sent a transaction or a message. There are many business-to-business (B2B) systems where non-repudiation is often an essential requirement. Considering the digital

signature is based on public key encryption, the sender of the message cannot repudiate successfully the fact that he signed the message. Although non-repudiation is tied in the context of an user signing something, we also can bring this term when discuss about an enterprise with portals, applications, and web services. A portal may sign a portion of its message to a Web service, and a Web service may sign a portion of its messages. A side effect of digitally signing a document is also integrity. Because the signed message is actually the signature of the hash of the message used for proving integrity, many simply view non-repudiation as very strong integrity. XML Signature is a W3C standard used for providing non-repudiation, and is used in other standards, such as the WS-Security standard. The following section describes such standards. [12]

2.6. Auditing

Audit is the process of verifying that security requirements have been satisfied, with corrections suggested where they haven't been met. Essential to effective auditing is that actions are traced and logged through all parts of the system. With web applications, this means that logging of significant operations must happen in the web server, web application and data layer, in addition to the web application itself. Events of interest include: errors, failures, state accesses, authentication, access control and other security checks, in addition to application-specific operations and actions. Care must be taken to protect the integrity of logging and trace data, even (and perhaps especially) in the case of system failures. Logs that are tampered with or destroyed are useless in performing an effective audit. Auditing of log and trace data can either be done manually or it can be automated and often a combination of both is used. Either way, auditing should be done on a regular basis. Automated systems that continually monitor, detect, and in some cases even correct (or at least recommend corrections for) security problems can be particularly useful for maintaining a secure web application. Somewhat related to logging and auditing, Web applications should be careful to ensure that errors or failures somewhere in the system do not introduce security vulnerabilities. Attackers, for instance, are often able to exploit the detailed error information provided by web applications to gain unauthorized access.

2.7. Session Management.

Web-based applications, in contrast to desktop-based application clients, have a challenge with regard to where client-related state information is stored. A desktop application would store state locally on the client machine, but because of the relatively stateless nature of the web browser, client state in web applications tends to be stored remotely on the server. The challenge then becomes securely managing and associating session state with an authenticated client identity. Unlike the other areas mentioned above, session state management is not strictly a security concern. However, the potential for security vulnerabilities in this area as well as its unique relevance to web applications merit its discussion here.

Many web application containers include built-in session management capabilities, and in most cases, it is desirable to leverage this functionality where possible.

When session state management must be built by the web application, care must be taken to ensure that session state can not be tampered with and is securely (i.e., cryptographically) and consistently mapped to an authentication token. From the client perspective, session identifiers are often included in cookies that are automatically saved and presented by the web browser. Such information could also be presented elsewhere in user input data. Care must be taken to protect the integrity and confidentiality of these session identifiers as attackers can use this

information to gain unauthorized access to the system (see the attack scenario below). As much as possible web application interfaces should be constructed such that users can keep their session state secure (often this means including sensible logout procedures, among other things).

- **Grid Portal Security Requirements**

So far, we discussed about security concepts (for portal). From now on, we will present the security needs of web applications discussed above.

To their advantage, Grid resources tend to have their own security (authentication and authorization in particular) mechanisms in place, so breaking into a Grid portal, while concerning, may not necessarily allow the attacker access to backend Grid resources. For instance, simply being able to submit jobs through a portal is not useful without proper Grid credentials to authenticate to the actual job execution service. Consequently, the key security challenge of most Grid portals is that at some level, *they manage Grid credentials on behalf of clients.* Compromised Grid credentials are an extremely serious security breach because they allow an attacker to effectively impersonate a valid Grid user until the credentials are revoked or expire. Thus, extra care must be taken in the management of these Grid credentials, which can effectively be viewed as a special kind of session state. The integrity and confidentiality of these credentials must be maintained even in the case of errors or failures. Accesses to the credentials should be logged and monitored continuously for suspicious behavior. Further the credentials, especially if stored on disk must be protected from other users or applications running on the web application server. A compromise elsewhere in the server's software stack should not lead to compromise of user's Grid credentials.

Vulnerabilities of Web Applications

A great challenge in developing secure web applications is that the vulnerabilities in any component of the architecture can often result in compromise of the web application as a whole. For instance, even though the code of a particular web application might be carefully written and free of security holes, vulnerabilities in the web server could still be exploited, causing the secure web application to be hijacked or overridden with a malicious version. Another challenge of the architectural complexity of many web applications is that it is often difficult to configure all of the components correctly and securely. So, even if the components as developed are free of security vulnerabilities, misconfiguration can unwittingly open the web application to compromise.

The Open Web Application Security Project (OWASP) compiled a list of ten of the most common security vulnerabilities afflicting Web applications [17] (and thus Grid portals, which are just a specific type of web application):

- **Unvalidated Parameters** – input contained in web requests is not properly checked (by the application) before being acted on. Attackers can craft parameters to hijack the application or cause it to behave in dangerous, unexpected ways. *Injection Flaws*, *Buffer Overflows*, and *XSS Flaws* are all specific types of *Unvalidated Parameter* vulnerabilities.
- **Broken Access Control** – access control mechanisms work inconsistently or incorrectly, allowing unintended access to resources. This is particularly troublesome for web application administrative interfaces.
- **Broken Authentication and Session Management** – authentication problems can range from weak authentication mechanisms that are easily broken (plain text secrets to retrieve forgotten passwords), to insufficient session protection (exploiting access to one set of session information to gain access to someone else's) to forged sessions or session cookies (allowing session impersonation).

- **Cross-Site Scripting (XSS) Flaws** – involves exploiting an invalidated parameter vulnerability to send a script to the web application that is in turn delivered to and executed by the end user's web browser. **Buffer Overflows** – specially crafted input results in the execution of arbitrary code on the target server. This is particularly problematic if the server is running as root or an administrator account as the malicious code will also have those privileges. In general, Java applications do not suffer from this type of vulnerability (although the JVM itself could).
- **Injection Flaws** – in contrast to the other invalidated parameter attacks, this refers to when injected code or command strings are passed through the web application directly to some backend system. SQL injection attacks are probably the most common.
- **Improper Error Handling** – this type of vulnerability surfaces when error messages displayed to the user in some way reveal details about how the system or application works (the attacker could then exploit this knowledge). This is typically a problem when very detailed stack traces are displayed to the user giving some information of the structure of the code and its operation. Attackers can also probe for inconsistencies in error messages returned ("file not found" vs. "access denied") to gain a better understanding of the application.
- **Insecure Storage** – storage of sensitive data (passwords, account information, etc.) without proper encryption or access control mechanisms. This could be on disk, in a database, or in memory. Usually one of the other exploits is needed to actually gain access to this insecure data.
- **Denial of Service** – when the sheer volume of requests to the web application overwhelms the capacity, denying access to legitimate users. This is usually an even more troublesome problem as web server DoS attacks (like SYN flooding), because it's very hard for web applications to distinguish between legitimate and malicious requests. The complexity of web applications usually means a fairly low threshold of concurrent connections needs to be exceeded to deny access.
- **Insecure Configuration Management** – problems here range from unpatched software to unchanged insecure default settings to outright configuration mistakes caused by incomplete or incorrect understanding of some very complex software. Clearly this is a human problem as much as a software problem, but delivering software that's easy to understand, easy to configure and comes in a secure configuration out of the box would certainly help.

2.8. Key Security Standards

With a significant role in portal development, we will continue with a brief overview of security standards which a portal developer needs to know.

2.8.1. SSL and TLS

Created by Netscape, SSL (Secure Sockets Layer) and TLS (Transport Layer Security) are higher-level encryption protocols that are used to assure data integrity and confidentiality between two points. Also, they can be used for mutual authentication when both parties have digital certificates. Although, the both are very similar (because TLS is based on SSL) and in most cases we simply refer to both protocols as SSL, they will not interoperate (there are subtle differences).

In an easy-going language, SSL with HTTP are called HTTPS Sessions (a process providing confidential web transmission). A portal developer consider that a SSL session can protect confidentiality and data integrity between the user and the portal, but also between the portal and its next communication point. [12]

2.8.2. XML Encryption

Being used to encrypt elements of an XML document, XML Encryption is a W3C

standard that operates XML documents and XML element-level confidentiality. It can be used with key exchange algorithms and public key encryption in order to encrypt documents to different parties. A great advantage of XML encryption (unlike SSL, which is decrypted at each point) is that it can be used in solutions with multiple network nodes between the portal and the data source.

2.8.3. XML Signature

XML Signature is a W3C standard that assures the message integrity and non-repudiation of XML documents. Any part of an XML document can be digitally signed – becoming self-validating when it has a public key. XML Signature is based on a public key technology in which the hash of a message is cryptographically signed; this provides integrity and non-repudiation. When portal communicates with Web services, XML Signature has an important role. Because of self-validating, user's credentials can be put on SOAP messages beyond the portal. [14]

2.8.4. SAML

Security Assertion Markup Language (SAML), is an OASIS standard which is used for pass authentication and authorization information between different parts. In a portal environment, a portal can "declare" that it authenticated a user, having in the same time some certain security credentials. A SAML assertion can be digitally signed using XML Signature. It is good to know that SAML can solve significant challenges in Web services security, because signed SAML can travel between different platforms and organizations. Anyone trusting the signer will trust the credential. SAML is an important standard, and many open-source toolkits are available. [12]

3. Grid construction – General principles

This section briefly highlights some of the general principles that underlie the construction of the Grid. In particular, the idealized design features that are required by a Grid to provide users with a seamless computing environment are discussed. Four main aspects characterize a Grid:

- **Multiple administrative domains and autonomy**. Grid resources are geographically distributed across multiple administrative domains and owned by different organizations. The autonomy of resource owners needs to be honored along with their local resource management and usage policies.
- **Heterogeneity**. A Grid involves a multiplicity of resources that are heterogeneous in nature and will encompass a vast range of technologies.
- **Scalability**. A Grid might grow from a few integrated resources to millions. This raises the problem of potential performance degradation as the size of Grids increases. Consequently, apps. that require a large number of geographically located resources must be designed to be latency and bandwidth tolerant.
- **Dynamicity or adaptability**. In a Grid, resource failure is the rule rather than the exception. In fact, with so many resources in a Grid, the probability of some resource failing is high. Resource managers or applications must tailor their behavior dynamically and use the available resources and services efficiently and effectively. [3]

The components that are necessary to form a Grid are as follows:

- **Grid fabric**. This consists of all the globally distributed resources that are accessible from anywhere on the Internet. These resources could be computers (such as PCs or Symmetric Multi- Processors) running a variety of operating systems (such as UNIX or Windows), storage devices, databases, and special scientific instruments such as a radio telescope or particular heat sensor.
- **Core Grid middleware**. This offers core services such as remote process management, co-allocation of resources,

storage access, information registration and discovery, security, and aspects of Quality of Service (QoS) such as resource reservation and trading.

- **User-level Grid middleware**. This includes application development environments, programming tools, and resource brokers for managing resources and scheduling application tasks for execution on global resources.
- **Grid applications and portals.** Grid applications are typically developed using Grid-enabled languages and utilities such as HPC++ or MPI. An example application, such as parameter simulation or a grand-challenge problem, would require computational power, access to remote data sets, and may need to interact with scientific instruments. Grid portals offer Web-enabled application services, where users can submit and collect results for their jobs on remote resources through the Web.

In attempting to facilitate the collaboration of multiple organizations running diverse autonomous heterogeneous resources, a number of basic principles should be followed so that the Grid environment:

- does not interfere with the existing site administration or autonomy;
- does not compromise existing security of users or remote sites;
- does not need to replace existing operating systems, network protocols, or services;
- allows remote sites to join or leave the environment whenever they choose;
- does not mandate the programming paradigms, languages, tools, or libraries that a user wants;
- provides a reliable and fault tolerant infrastructure with no single point of failure;
- provides support for heterogeneous components;
- uses standards, and existing technologies, and is able to interact with legacy applications;
- provides appropriate synchronization and component program linkage. [3]

As one would expect, a Grid environment must be able to interoperate with a whole spectrum of current and emerging hardware and software technologies. An obvious analogy is the Web. Users of the Web do not care if the server they are accessing is on a UNIX or Windows platform. From the client browser's point of view, they 'just' want their requests to Web services handled quickly and efficiently. In the same way, a user of a Grid does not want to be bothered with details of its underlying hardware and software infrastructure. A user is really only interested in submitting their application to the appropriate resources and getting correct results back in a timely fashion. An ideal Grid environment will therefore provide access to the available resources in a seamless manner such that physical discontinuities, such as the differences between platforms, network protocols, and administrative boundaries become completely transparent. In essence, the Grid middleware turns a radically heterogeneous environment into a virtual homogeneous one.

The following are the main design features required by a Grid environment.

- Administrative hierarchy. An administrative hierarchy is the way that each Grid environment divides itself up to cope with a potentially global extent. The administrative hierarchy determines how administrative information flows through the Grid.
- Communication services. The communication needs of applications using a Grid environment are diverse, ranging from reliable point-to-point to unreliable multicast communications.

The communications infrastructure needs to support protocols that are used for bulk-data transport, streaming data, group communications, and those used by distributed objects.

- The network services used also provide the Grid with important QoS parameters such as latency, bandwidth, reliability, fault-tolerance, and jitter control.
- Information services. A Grid is a dynamic environment where the location and types of services available are constantly changing. A major goal is to make all resources accessible to any process in the system, without regard to the relative location of the resource user. It is necessary to provide mechanisms to enable a rich environment in which information is readily obtained by requesting services. The Grid information (registration and directory) services components provide the mechanisms for registering and obtaining information about the Grid structure, resources, services, and status.
- Naming services. In a Grid, like in any distributed system, names are used to refer to a wide variety of objects such as computers, services, or data objects. The naming service provides a uniform name space across the complete Grid environment. Typical naming services are provided by the international X.500 naming scheme or DNS, the Internet's scheme.
- Distributed file systems and caching. Distributed applications, more often than not, require access to files distributed among many servers. A distributed file system is therefore a key component in a distributed system. From an applications point of view it is important that a distributed file system can provide a uniform global namespace, support a range of file I/O protocols, require little or no program modification, and provide means that enable performance optimizations to be implemented, such as the usage of caches. [3]
- Security and authorization. Any distributed system involves all four aspects of security: confidentiality, integrity, authentication, and accountability. Security within a Grid environment is a complex issue requiring diverse resources autonomously administered to interact in a manner that does not impact the usability of the resources or introduces security holes/lapses in individual systems or the environments as a whole. A security infrastructure is the key to the success or failure of a Grid environment.
- System status and fault tolerance. To provide a reliable and robust environment it is important that a means of monitoring resources and applications is provided. To accomplish this task, tools that monitor resources and application need to be deployed.
- Resource management and scheduling. The management of processor time, memory, network, storage, and other components in a Grid is clearly very important. The overall aim is to efficiently and effectively schedule the applications that need to utilize the available resources in the Grid computing environment. From a user's point of view, resource management and scheduling should be transparent; their interaction with it being confined to a manipulating mechanism for submitting their application. It is important in a Grid that a resource management and scheduling service can interact with those that may be installed locally.
- Computational economy and resource trading. As a Grid is constructed by coupling resources distributed across various organizations and administrative domains that may be owned by different organizations, it is essential to support mechanisms and policies that help in regulate resource supply and demand [1], [2]. An economic approach is one means of managing resources in a complex and decentralized manner. This approach provides incentives for resource owners, and users to be part of the Grid and develop and using strategies that help maximize their objectives.
- Programming tools and paradigms. Grid applications (multi-disciplinary apps.) couple resources that cannot be replicated at a single site even or may be globally located for other practical reasons. A Grid should include interfaces, APIs, utilities,

and tools to provide a rich development environment. Common scientific languages such as C, C++, and Fortran should be available, as should application-level interfaces such as MPI and PVM. A variety of programming paradigms should be supported, such as message passing or distributed shared memory. In addition, a suite of numerical and other commonly used libraries should be available.

- User and administrative GUI. The interfaces to the services and resources available should be intuitive and easy to use. In addition, they should work on a range of different platforms and operating systems. They also need to take advantage of Web technologies to offer a view of portal supercomputing. The Web-centric approach to access supercomputing resources should enable users to access any resource from anywhere over any platform at any time. That means, the users should be allowed to submit their jobs to computational resources through a Web interface from any of the accessible platforms such as PCs, laptops, or Personal Digital Assistant, thus supporting the ubiquitous access to the Grid. The provision of access to scientific applications through the Web (e.g. RWCPs parallel protein information analysis system [16]) leads to the creation of science portals. [3]

4. Conclusion

Nowadays, security is a hot topic and it is obvious that data needs to be protected. Taking into consideration that architects, portal administrators and, of course, developers are faced with a variety of factors when planning for portal application security, it is very important do not forget the significant role of security aspects and required standards in portal development.

According to the American Heritage Dictionary, a portal is "a doorway, entrance, or gate, especially one that is large and imposing". The intent behind such structures is really one of security, to allow the welcome visitors through, while keeping unwelcome intruders out. From a technological perspective, a portal is something that provides a convenient entry point to resources, applications or content located elsewhere. Early *Web portals* were typically web sites with search engines or indexes to other content on the World Wide Web.

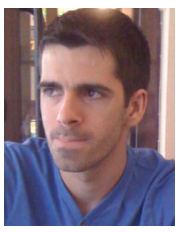
Since all of the content accessible through these web portals was publicly available anyway, everyone was welcomed in and security was barely a concern. In Grid computing, the resources of interest are not websites, but data and computational resources, services and applications. Thus, the goal of a *Grid portal* is to provide a convenient entry point to these Grid resources, typically via a Web-based front-end. While many Grid portals expose relatively general purpose functionality like launching jobs for remote execution or retrieving remotely-stored data, they can also include application specific interfaces customized for a particular domain.

Security gains prominence in Grid portals largely because of the nature of the Grid resources they expose. Many Grids link together powerful clusters of computational power and large scale data stores containing confidential, classified or proprietary information. A compromised Grid portal could allow an attacker to harness these powerful computational resources to launch a large scale attack elsewhere on the Internet or to gain user access to probe for privilege escalation or root compromise, for example.

Generally speaking, "*We've done tremendous work to secure computers but nothing to secure the human operating system. To change human behaviour, you need to educate and train employees, not just once a year but continuously. Like you continually patch computers and applications, you're continually training and patching human operating systems.*" ([5], pp.1)

References

[1] Buyya R, Abramson D, Giddy J. Economy driven resource management architecture for computational power grids. *International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'2010)*, Las Vegas, NV, 2010. CSREA Press: Athens, GA
[2] Buyya R. Economic-based distributed resource management and scheduling for Grid computing. *PhD Thesis*, Monash University, Melbourne, Australia, April 2010.
[3] David Del Vecchio, Victor Hazlewood and Marty Humphrey, "Evaluating Grid Portal Security" *Department of Computer Science, University of Virginia*, San Diego, 2009
[4] Hazen A.Weber, "Role-Based Access Control: The NIST Solution", *SANS Inst.*, 08
[5] Lance Spitzner, "Target: The Human", *Information Security Magazine*, May 2011
[6] Mark Baker, Rajkumar Buyya and Domenico Laforenza, "Grids and Grid technologies for wide-area distributed computing", *School of Computer Science, University of Portsmouth,* Mercantile House, Portsmouth, U.K. and *Grid Computing and Distributed Systems Laboratory, Department of Computer Science and Software Engineering*, The University of Melbourne, Melbourne, Australia
[7] M. Velicanu, D. Litan, L. Copcea (Teohari), M. Teohari, A.M. Mocanu (Virgolici), I. Surugiu, O. Raduta, "Ways to Increase the Efficiency of Information Systems", *The Proc. of the 10th WSEAS Internat. Conf. on artificial Intelligence, Knowledge Engineering and Databases, Cambridge, UK*, 2011.
[8] M. Velicanu, D. Litan, I. Surugiu, O. Raduta, A.M. Mocanu (Virgolici), "Information Technology Standards – a Viable Solution to Reach the Performance", *International Conference on TECHNOLOGY POLICY and LAW (TPL '11), Brasov, RO*, 2011.
[9] D. Litan, L. Copcea (Teohari), M. Teohari, A.M. Mocanu (Virgolici), I. Surugiu, O. Raduta, "Information Systems Integration, a New Trend in Business", *APPLICATIONS of COMPUTER ENGINEERING (ACE '11), Canary Islands, ES*, 2011.
[10] National Institute of Standards and Technology, Computer Security Resource Center, "An Introduction to Role-Based Access Control, in *ITL Computer Security Bulletin.*, Dec. 2010.
[11] Ovidiu Raduta, Adrian Munteanu, "Business Intelligence Solutions - Security Components", *The 10th International Conference on Informatics in Economy*, 2011
[12] W. Clay Richardson, Donald Avondolio, Joe Vitale, Peter Len, Kevin T. Smith, "Professional Portal Development with Open Source Tools: Java™ Portlet API, Lucene, James, Slide", *Wiley Technology Publishing*, 2009
[13] http://www.sans.org
[14] www.rsa.com
[15] http://csrc.nist.gov/publications
[16] http://www.rwcp.or.jp/papia/
[17]http://www.owasp.org/documentation/topten.html
[18]http://en.wikipedia.org/wiki/Web_portal



Ovidiu Rădută has graduated the Academy of Economic Studies (Bucharest, Romania), Faculty of Cybernetics, Statistics and Economic Informatics in 2006. He holds a Master diploma in Informatics Security (Master Thesis: IT Software in banks. Security Issues) from 2008 and currently, he is a Ph.D. Candidate in Economic Informatics with his Doctor's Degree Thesis: Bank System's Process Optimizing. In present, he is ISTQB – Advanced Test Analyst certified and he works as Senior Test Analyst with 3+ years testing experience in Raiffeisen Bank Romania (6+ years banking projects experience). His research

activity can be observed in many international proceedings (papers ISI proceedings). His scientific fields of interest include: Test management, Test Techniques, Databases processes, Middleware Products, Information Systems and Economics.

Adrian Munteanu has graduated the Academy of Economic Studies (Bucharest, Romania), Cybernetics, Statistics and Economic Informatics in 2001. Currently, he is a Ph.D. Candidate in Economic Informatics with his Doctor's Degree Thesis: DataWarehouses - Business Support. In present, he is Advanced Resolution Engineer with 12+ years experience in database and Enterprise solutions field at Oracle Corporation. His research activity can be observed in many international proceedings (papers ISI proceedings) published by now. His scientific fields of interest include: Business Intelligence, Datawarehouse Modelling and Enterprise Resource Planning implementation.