THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

DATABASE SYSTEMS JOURNAL

Vol. V, Issue 1/2014

LISTED IN

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

BUSINESS INTELLIGENCE

ERP

DATA MINING

DATA WAREHOUSE

DATABASE

ISSN: 2069 – 3230

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

CONTENTS

New Classes of Applications in the Cloud. Evaluating Advantages and Disadvantages of Cloud Computing for Telemetry Applications

Anca APOSTU, Florina PUICAN, Geanina ULARU,
George SUCIU, Gyorgy TODORAN
[1, 2, 3]University of Economic Studies, Bucharest, Romania
[4, 5]University Politehnica of Bucharest, Romania
ancaiapostu@gmail.com, puicanflorina@yahoo.com, ularugeanina@yahoo.com,
george@beia.ro, todoran.gyorgy@gmail.com

Nowadays companies are moving some parts of their businesses to the cloud. Industry predictions are that this trend will continue to grow and develop even further in the coming few years. While Cloud computing is undoubtedly beneficial for mid-size to large companies, it is not without its downsides, especially for smaller businesses. This paper's aim is to deliver an analysis based on advantages and disadvantages of Cloud computing technology, in order to help organizations fully understand and adopt this new computing technology. Finally, to prove the advantages that Cloud technology can have for this domain, we are presenting a cloud application for telemetry, with a focus on monitoring hydro-energy. We consider that the way to attain the benefits of Cloud technology is to understand its strengths and weaknesses and adapt to them accordingly.

Keywords: *Advantages, Architecture, Cloud Computing, Grid Computing, Telemetry*

1 Introduction

Nowadays evolution has its premises on the fact that faster access to innovation drives higher productivity. The Web is recognized as epicenter of innovation. Rapid innovation powered by the Cloud has an advantage over traditional technology cycles: employees adapt to a continuous stream of manageable improvements better than they tolerate large, disruptive batches of change. Gradual iterations in bite-sized chunks substantially reduce change-management challenges. Conversely, employees are subjected to a painful re-learning cycle each time companies upgrade traditional software.

2. Cloud computing - actual context

Because data is stored in the Cloud instead of on employee computers, Cloud computing enhances multiple users to access and contribute to projects simultaneously without worrying about using the same operating system, software, or browser. For example, instead of collaborating on a document by sending back and forth revision after revision as attachments, documents are stored in the cloud. Co-workers can access the web-based document simultaneously in their browsers, and even make changes that other authorized users can see in real-time. Eliminating attachment round-trips by storing data in the cloud saves time and reduces frustrations for teams who need to work together efficiently.

Through synchronous replication, data and user actions are mirrored in nearly real-time across multiple data centers. If one data center becomes unavailable for any reason, the system is designed to instantly fall back to a secondary data center with no user-visible interruption in service.

Cloud provides extensive flexibility and control. Nevertheless, moving to the cloud doesn't mean that businesses lose control of their data or their technology. For example, the Google Apps Terms of Service explicitly state that customers retain ownership of their data in Google Apps.

Furthermore, cloud providers give controls so administrators can manage which applications their users can access and how employees can use each service. They also allow administrators build custom functionality and integrations with other technologies.

Going detailed in the topic of Cloud Computing we must mention that Cloud Computing is split in three different categories according to [1]:

1. *IaaS - Infrastructure as a Service*: Virtual provision of computing power and/or memory. Source [2] mentions as prominent example of an IaaS service the Amazon WS service.
2. *PaaS – Platform as a Service:* Provision of a runtime environment, like application servers, databases. In this area, paper [2] provides Google's App Engine as probably the most prominent example.
3. *SaaS – Software as a Service:* Provision of usually browser based applications that can directly be used. Google Docs or the Customer Relationship Management software of salesforce.com might serve as examples.

3. Evolution is shown by new achievements - distributed computing versus grid computing

Cloud evolved from Grid computing, but the latter can function separately. Cloud definition usually superposes with Grid computing technology or more generally with distributed computing definition. From the end user perspective, the interest in what actually happens "behind the scenes" in Cloud is minimum, in comparison with system administrators who virtualize servers and handle applications in cloud. Grid Computing is the infrastructure on which cloud computing relies.

There are differences and similarities between the two mentioned technologies. Cloud and grid computing assure scalability, they are multitasking and share resources among a large number of end users. The differences come by analyzing the computing model, data management, the visualization or security model. Grid Computing "enables resource sharing and coordinated problem solving in dynamic, virtual organizations" [3]. From computing model perspective, grid computing uses batch computation and via batches there are identified users and the number of processors required, whereas Cloud computing functions with resources shared by users in the same time.

Data management structure is very important to provide management implementation to the needed data and also fast and efficient data retrieval. Grid computing is using data ware schedulers [4], but Cloud might be challenged by the data handling from applications, without investing in the data access patters. Virtualization and encapsulation are very used in cloud and more intensively in grid computing, because the grid holds the control on the resources, without necessarily virtualizing them.

From the security model perspective, there might be a potential issue in cloud. For data protection, the users might desire to manage their own private keys, but for this, detailed private key management should be provided. Nevertheless, from Grid computing perspective resources are heterogeneous and have their autonomy. The security in Grid computing is assured in the infrastructure.

Our comparison between the two technologies puts in spot light the common share of visions and architectures, but also the differences between them at the data management and security model. We have identified the weaknesses that should be overcome by both technologies in order to speed up their evolution.

4. Analysis of advantages and disadvantages of cloud computing

In the following section we are presenting the main advantages and disadvantages of Cloud Computing applying them for telemetry applications.

A. Advantages of Cloud Computing

Speaking about advantages of Cloud Computing we present bellow the main benefits for businesses in general, focusing at some points on examples for small businesses:

- *Cost efficiency* - Cloud computing is probably the most cost efficient method to use, maintain and upgrade, as explained in [5]. Traditional desktop software costs companies a lot, in terms of finance. Adding up the licensing fees for multiple users can prove to be very expensive for the establishment concerned. The cloud, on the other hand, is available at much cheaper rates and hence, can significantly lower the company's IT expenses. Besides, there are many one-time-payments, pay-as-you-go and other scalable options available, which make it very reasonable for the company in question. Paper [6] adds up that it lowers the cost for smaller firms which intend to apply the compute-intensive techniques.
- *Almost Unlimited Storage.* Storing information in the cloud gives you almost unlimited storage capacity.
- *Backup and Recovery.* Since all the data is stored in the cloud, backing it up and restoring the same is relatively much easier than storing the same on a physical device. Furthermore, most cloud service providers are usually competent enough to handle recovery of information. Hence, this makes the entire process of backup and recovery much simpler than other traditional methods of data storage.
- *Automatic Software Integration.* In the cloud, software integration is usually something that occurs automatically. This means that Cloud users don't need to take additional efforts to customize and integrate their applications as per own preferences. This aspect usually takes care of itself.
- *Easy Access to Information.* Once the users register in the cloud, they can access the information from anywhere, where there is an Internet connection. This convenient feature lets users move beyond time zone and geographic location issues.
- *Quick Deployment.* Lastly and most importantly, Cloud computing gives the advantage of quick deployment. Once opting for this method of functioning, the entire system can be fully functional in a matter of a few minutes. Of course, the amount of time taken here will depend on the exact kind of technology that is needed for the business.
- *Easier scale of services.* It makes it easier for enterprises to scale their service according to the demand of clients.
- *Deliver new services.* It makes possible new classes of applications and deliveries of new services that are interactive in nature.

B. Performance achievement with Cloud Technologies and Parallel Computing

Among the benefits of Cloud Computing there can be mentioned the accessibility to customized virtual machines, the payment done for what it is used and efficient resource allocation. Cloud computing brings advantages not only to large companies, but also to small and medium-sized ones, by outsourcing data infrastructure. The data can be accessed from any location, from the clouds.

Better performance is achieved in the context of parallel computing with Cloud technologies. Applications that encounter latencies can overcome their deficiencies by utilizing technologies such as Apache Hadoop (a study on Apache Hadoop is presented in paper [7]), MapReduce (former CGL-MapReduce) and Dryad. Nevertheless, more complex applications, with higher expectations from the performance point of view, require communication paradigms and customized network settings such as MPI (Message

Passing Interface), a standardized API used to implement parallel applications.

The MPI implications for virtualized resources might be analyzed through its implementation. The analysis of performance achievement implies the understanding of the complex process of the application's adoption of MPI and its impact on cloud resources. MPI sustains I/O operations, collective communication and point-to-point communication [6]. The improvements of MPI on the application reflect the mapping of the processors from the clusters. The CPUs evaluation might provide indicators regarding this aspect.

From the performance point of view we propose a comparison for different characteristics of the parallel computing technologies. From the programming languages perspective, for MPI, there are used C++, Java and C#, for Dryad there are C# and DryadLINQ, for MapReduce and Hadoop the main used language is Java. The data usage is assured by MPI, Dryad and MapReduce through directories, shared files and local disks and for Hadoop by HDFS. The communication is achieved in MapReduce by distribution network, in Hadoop by HDFS, in MPI and Dryad by files and TCP pipes. The failures are worked on differently according to what technology is used; for MPI there is OpenMPI and for Dryad the failure is handled by the re-execution of maps.

Cloud technologies enhance the way Big Data is handled and the processes used for failures approaches. The minuses might be considered when the computation is moved to data and the parallel computing is done on the local storage.

C. Disadvantages of Cloud Computing

In spite of its many benefits, as mentioned above, Cloud computing also has its disadvantages. Businesses, especially smaller ones, need to be aware of these aspects before going in for this technology. The main risks involved in Cloud Computing are:

- *Technical Issues.* Though it is true that information and data on the Cloud can be accessed any time and from anywhere, there are moments when the system can have some serious malfunction. Businesses should be aware of the fact that this technology is always prone to outages and other technical issues. Even the best Cloud service providers run into this kind of trouble, in spite of keeping up high standards of maintenance.
- *Security in the Cloud.* The other major issue of Cloud is represented by security. Before adopting this technology, beneficiaries should know that they will be surrendering all their company's sensitive information to a third-party cloud service provider. This could potentially impose a great risk to the company. Hence, businesses need to make sure that they choose the most reliable service provider, who will keep their information totally secure. "Switching to the cloud can actually improve security for a small business", as mentioned by Michael Redding, managing director of Accenture Technology Labs, cited by [8]. "Because large cloud computing companies have more resources, he says, they are often able to offer levels of security an average small business may not be able to afford implementing on its own servers" [8].
- *Prone to attack.* Storing information in the cloud could make the companies vulnerable to external hack attacks and threats; therefore there is always the lurking possibility of stealth of sensitive data.
- *Possible downtime.* Cloud computing makes the small business dependent on the reliability of their Internet connection.
- *Cost.* At first glance, a cloud computing application may appear to be a lot cheaper than a particular software solution installed and run in-house. Still, the companies need to

ensure that the cloud applications have all the features that the software does and if not, to identify which are the missing features important to them. A total cost comparison is also required. While many cloud computer vendors present themselves as utility-based providers, claiming that they only charge for what customers use, Gartner says that this isn't true; in most cases, a company must commit to a predetermined contract independent of actual use. Companies need to look closely at the pricing plans and details for each application.

Furthermore, "the increase of the demand for computing resources has led to the deployment of cloud computing data centers and to a corresponding and significant increase of the total energy consumed by these infrastructures", as paper [9] explains.

- *Inflexibility*. Choosing a Cloud computing vendor often means locking the business into using their proprietary applications or formats. For instance, it is not possible to insert a document created in another application into a Google Docs spreadsheet. Furthermore, a company needs to be able to add and/or subtract Cloud computing users as necessary as its business grows or contracts.
- *Lack of support.* Anita Campbell (OPEN Forum) writes, "Customer service for Web apps leaves a lot to be desired - all too many cloud-based applications make it difficult to get customer service promptly – or at all. Sending an email and hoping for a response within 48 hours is not an acceptable way for most of us to run a business" [10]. The New York Times writes: "The bottom line: If you need handholding or if you are not comfortable trying to find advice on user forums, the cloud probably is not ideal" [11].

Paper [12] adds up "some of the major technical risks, which include: multi-tenant environment; internet as connection; system complexity and loss of control." Also "new vulnerabilities inherent to Cloud computing include breaches from one virtual computing space to another, misappropriation of session security from web protocols, and limited encryption capabilities in many protocols."

As paper [13] explains about adoption of Cloud computing, "it doesn't mean that every small business should immediately throw out all their servers and software and conduct all their business operations in the cloud". Small business owners have different needs and different comfort levels. It may be more advantageous for you to use cloud computing only for certain applications. Or even not at all. Previously to adopting Cloud computing, business owners should consider how these disadvantages of cloud computing could affect their small business.

A very interesting comparative study based on cloud Computing is presented in paper [14] following a series of surveys conducted in Australia and Czech Republic at the end of 2011. The aim of this study was "to identify differences between adoption patterns in countries with different level of readiness for Cloud computing". An interesting result of this research relates to changes in perception of adoption issues following a decision to adopt cloud services. The results indicate that "concerns about data security, IT governance, service availability and dependence on service provider are held by much lower number of respondents following adoption of cloud services". Finally, the study indicates that "SaaS adoption is confined to a small number of relatively simple types of enterprise applications that include CRM, email and other types of collaboration software".

One can only say that Cloud computing is a complex and rapidly evolving concept, therefore companies should consider some important aspects when planning their initiative towards adopting this technology:

- ✓ Understand what Cloud computing technology is, how it will evolve and under what circumstances it can offer value;
- ✓ Evaluate models, architectures, technologies and IT organization best practices which are suitable for the companies which want to adopt Cloud computing in order to build private cloud computing environments.
- ✓ Consider how IT will secure, manage and govern cloud services across public, community, private and hybrid environments.
- ✓ Determine the which possibilities are for migrating applications to the cloud and if this brings value to the company and also determine opportunities to create “new cloud-optimized applications”[15].
- ✓ Analyze the way in which Cloud computing will affect the strategy and direction of IT and identify the opportunities for the enterprise to “provide cloud services to customers or partners”.

D. Cloud Advantages for Telemetry Applications

After analyzing the advantages and disadvantages of Cloud, in this chapter we present a Cloud test platform for clean energy production telemetry, with focus on hydro-energy. We use different types of RTU’s (Remote Telemetry Units) and sensors that monitor and transmit important information from selected locations such as temperature, precipitation, water level in the dam, quantity of water captured during winter or summer.

The RTUs transmit sensor data over GSM/GPRS to our cloud platform where we can conveniently process the site specific data in near real-time, display it in our web-based visualization application and get detailed recommendations when and where to generate energy - resulting in optimized energy production and income. Our system can be connected with other management systems to make better use of resources keeping in view other factors like energy price, consumption trends and to improve risk management [16].

Variable prices and rising costs of production are forcing energy producers to optimize production costs. Therefore "precision energy production", the optimized use of natural energy resources such as water, sun or wind is now indispensable. The growing environmental awareness of consumers further accelerates this process and promotes the usage of remote automatic monitoring system for field information such as the one we developed.

We will introduce SlapOS [17], the first open source operating system for Distributed Cloud Computing. SlapOS is based on a grid computing daemon called slapgrid which is capable of installing any software on a PC and instantiate any number of processes of potentially infinite duration of any installed software. Slapgrid daemon receives requests from a central scheduler the SlapOS Master which collects back accounting information from each process.

SlapOS[17] is an open source Cloud Operating system which was inspired by recent research in Grid Computing and in particular by Bonjour Grid [18], a meta Desktop Grid middleware for the coordination of multiple instances of Desktop Grid middleware. It is based on the motto that ”everything is a process”.

SlapOS Master follows an Enterprise Resource Planning (ERP) model to handle at the same time process allocation optimization and billing. SLAP stands for “Simple Language for Accounting and Provisioning”.

This structure has been implemented for cloud-based automation of ERP and CRM software for small businesses and aspects are under development under the framework of the European research project “Cloud Consulting”. We will use our platform hosted on several servers running Ubuntu Linux – Apache – MySQL template with current software release.

On our cloud testing environment we provide the platform for processing information from hundreds different sensors, enabling the analysis of environmental data through a large sample of RTUs. In previous approaches RTUs were implemented in most cases on a local server and no company could aggregate enough sensor data to consider automating the production process and providing the required resilience [19].

E. Cloud Telemetry Components

Telemetry systems have a large area of usage, as presented for example in paper [20], for "rain rate measurements" or for "distributed wells" in paper [21].

1) Cloud Architecture

SlapOS is based on a Master and Slave design. Slave nodes request to Master nodes which software they should install, which software they show run and report to Master node how much resources each running software has been using for a certain period of time. Master nodes keep track of available slave node capacity and available software. Master node also acts as a Web portal and Web service so that end users and software bots can request software instances which are instantiated and run on Slave nodes.

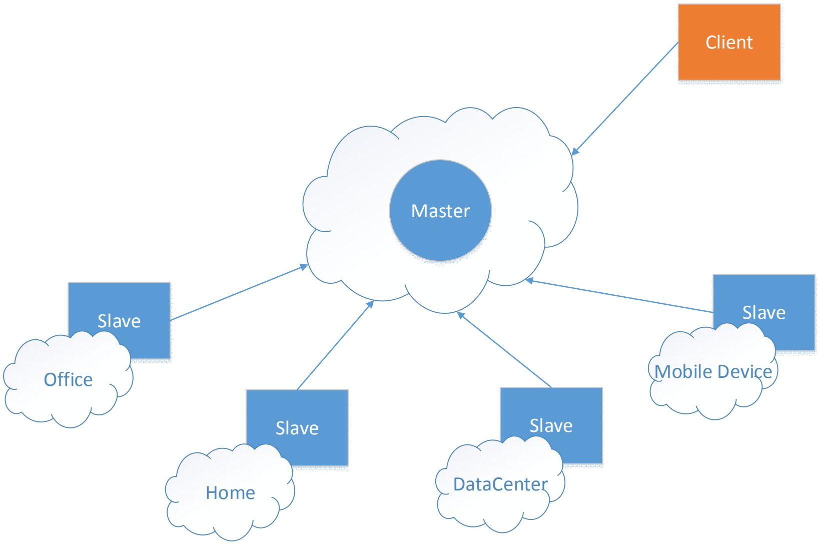


Fig. 1 SlapOS Master – Slave Cloud Architecture

Master nodes are stateful. Slave nodes are stateless. More precisely, all information required to rebuild a Slave node is stored in the Master node. This may include the URL of a backup service which keeps an online copy of data so that in case of failure of a Slave node, a replacement Slave node can be rebuilt with the same data.

It is thus very important to make sure that the state data present in Master node is well protected. This could be implemented by hosting Master node on a trusted IaaS infrastructure with redundant resource. Or - better - by hosting multiple Master nodes on many Slave nodes located in different regions of the world thanks to appropriate data redundancy heuristic. We are touching here the first reflexive nature of SlapOS. A SlapOS master is normally a running instance of SlapOS Master software instantiated on a collection of Slave nodes which, together, form a trusted hosting infrastructure. In other terms, SlapOS is self-hosted.

SlapOS Slave nodes are relatively simple compared to the Master node. Every slave node needs to run software requested by the Master node. It is thus on the Slave nodes that software is installed. To save disk space, Slave nodes only install the software which they really need.

Each slave node is divided into a certain number of so-called computer partitions. One may view a computer partition as a lightweight secure container, based on Unix users and directories rather than on virtualization. A typical bare-bone PC can easily provide 100 computer partitions and can thus run 100 RTU web portals or 100 sensors monitoring sites, each of which with its own independent database. A larger server can contain 200 to 500 computer partitions.

2) Telemetry Architecture

In Fig. 1 we present the general structure of the system that we are proposing for the tele-monitoring of installation sites in hydro power stations. At each of the monitored installation site is mounted an installation built mainly from distant RTU, sensors and actuators. There will be used especially RTUs capable to communicate with the Gateway through GSM-GPRS and Internet. For the installation sites which are situated in no GSM coverage areas will be used RTUs in the UHF band of 430-440 MHz. These will communicate with the date concentrator through a bridge station (bridge) which will ensure the UHF-GPRS and GPRS-UHF conversion. In the relatively few instances when this will be possible, the RTU-Gateway communication will be held radio exclusively in the UHF band of 430-440 MHz.

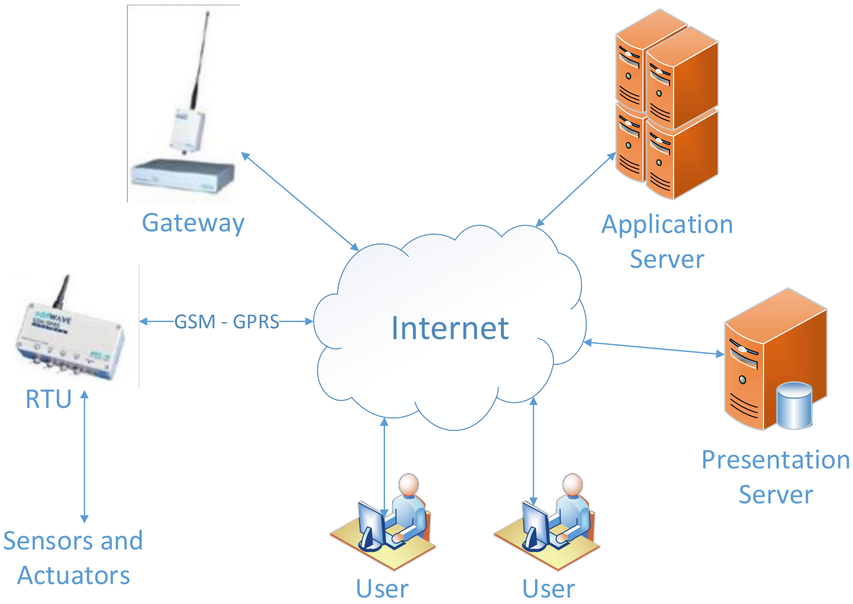


Fig. 2 General structure of the tele-monitoring system

The key elements of the system are:

- ✓ *Gateway*, which ensures the communication with the RTUs and available resource management;
- ✓ *Presentation Server (PS)* which is hosted on a computer with server features (for example, unattended operation 24/24), equipped with a software packet focused mainly on data presentation in various forms, entirely available to users.
- ✓ *Application Server (AS)*, focused on special tasks, which PS can't perform.

Practically, all system communication is done through Internet and this gives the system investment and mostly operational advantages. It is mentioned that the users can access the processed data, offered by the PS and AS anywhere and anytime, from any terminal with Internet access (PC, tablet, mobile phone etc.). The system's central elements are configured and scaled so that they would allow a system takeover of 100 RTUs

5. Conclusions

We conclude that while Cloud computing technology can prove to be a great asset to companies, it could also cause harm if not understood and used properly.

We consider Cloud computing to be an opportunity for small businesses to balance the efforts implied by IT management of course limited by the disadvantages of Cloud, some of them presented in this paper. The first and most important concern is given by security issues related to having their business data in the Cloud or, in a simpler way, having their data out on the Internet. Nevertheless, the recommendation would be to begin adopting Cloud Computing for a smaller part of their business applications in order to be able to count down the benefits and also to identify the risks.

6. Future researches

"Companies that are dependent on the IT environment of today need to remain competitive and, in order to do this, they need to keep up with the most recent technologies such as Cloud computing, mobile devices and virtualization" [22]. As identified by Gartner's Symposium/ITxpo in Orlando 2012 [15], Personal Cloud, Hybrid IT & Cloud Computing and Big Data will be between the most important ten strategic technology trends for 2013. Except for the much debated advantages of Cloud Computing these three trends represent major Cloud advances in the future and these will be subject for our future research.

References

[1] Chappell, D., *A short introduction to cloud platforms: An enterprise-oriented view,* White Paper, 13 pages, San Francisco, Chappell and Associates, 2008

[2] Marc Jansen, What does it service management look like in the Cloud? An ITIL based approach, *Proceedings of the International Conference on Computers, Digital Communications and Computing (ICDCC'11),* Barcelona, Spain, September 15-17, 2011, pp. 87-92, ISBN: 978-1-61804-030-5

[3] I. Foster, C. Kesselman, J. Nick, S. Tuecke, *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*. Globus Project, 2002.

[4] I. Raicu, Y. Zhao, I. Foster, A. Szalay. Accelerating Largescale Data Exploration through Data Diffusion, *International Workshop on Data-Aware Distributed Computing,* 2008.

[5] P. Viswanathan, *Cloud Computing – Is it Really All That Beneficial? Advantages and Disadvantages of Cloud Computing,* http://mobiledevices.about.com/od/additionalresources/a/Cloud-Computing-Is-It-Really-All-That-Beneficial.htm

[6] C. Evangelinos, C. Hill., Cloud Computing for parallel Scientific HPC Applications: Feasibility of Running Coupled Atmosphere-Ocean Climate Models on Amazon's EC2, *The First Workshop on Cloud Computing and its Applications* (CCA'08), 2008, Chicago

[7] E. G. Ularu, F. C. Puican, A. Apostu, M. Velicanu, *Perspectives on Big Data and Big Data Analytics*, Database Systems Journal vol. III, no. 4/2012, pp.3-14, ISSN: 2069 – 3230

[8] O. EL Akkad, *Outsource IT Headaches to the Cloud*, The Globe and Mail, 11 November 2010.

[9] J. Rietz, R. Macedo, C. Alves, J. V. De Carvalho, *Efficient Lower Bounding Procedures with Application in the Allocation of Virtual Machines to Data Centers,* WSEAS Transactions on Information Science & Applications, Volume 8, 2011, Print ISSN: 1790-0832, E-ISSN: 2224-3402, Available: http://www.wseas.us/e-library/transactions/information/2011/53-343.pdf

[10] A. Campbell, *These Issues Need to be Resolved Before Cloud*

Computing Becomes Ubiquitous, 11 August 2010, https://www.openforum.com/articles/these-issues-need-to-be-resolved-before-cloud-computing-becomes-ubiquitous-1/

[11] D. H. Freedman, *Thinking about moving to the Cloud? There are trade-offs*, 21 September 2011, The New York Times

[12] J.L. Kourik, *For Small and Medium Size Enterprises (SME) Deliberating Cloud Computing: A Proposed Approach,* Proceedings of the European Computing Conference (ECC '11) Paris, France, April 28-30, 2011, Available: http://www.wseas.us/e-library/conferences/2011/Paris/ECC/ECC-35.pdf

[13] S. Ward, *5 Disadvantages of Cloud Computing. Consider These Before You Put Your Small Business In the Cloud*, http://sbinfocanada.about.com/od/itmanagement/a/Cloud-Computing-Disadvantages.htm

[14] G. Feuerlicht, N. Margaris*, Cloud Computing Adoption: A comparative study,* Proceedings of the 1st WSEAS International Conference on Cloud Computing (CLC '12), Vienna, Austria, November 10-12, 2012, Available: http://www.wseas.us/e-library/conferences/2012/Vienna/COMPUTERS/COMPUTERS-71.pdf

[15] Press Release, Gartner Identifies the Top 10 Strategic Technology Trends for 2013, http://www.gartner.com/newsroom/id/2209615

[16] Z. Bocheng, "Design of Building Energy Monitoring and Management System," Second *International Conference on Business Computing and Global Informatization (BCGIN),* pp.645-648, Oct. 2012.

[17] J.P. Smets-Solanes, C. Cerin, and R. Courteaud, "SlapOS: A Multi-Purpose Distributed Cloud Operating System Based on an ERP Billing Model," *IEEE International Conference on Services Computing (SCC)*, pp.765-766, July 2011.

[18] H. Abbes, C. Cerin, and M. Jemni, *"A decentralized and fault-tolerant Desktop Grid system for distributed applications"* Concurrency and Computation: Practice and Experience volume 22, issue 3, pp. 261-277 2010.

[19] G.Suciu, C. Cernat, G. Todoran, G. Suciu, V. Poenaru, T. Militaru, and S. Halunga, "A solution for implementing resilience in open source Cloud platforms," *9th International Conference on Communications (COMM),* pp.335-338, June 2012.

[20] D. Kandris, G. Kaltsas, C. Nomicos, *Evaluation of Rain Rate Measurement Methods, Sensors and Systems,* WSEAS / IASME International Conference on Energy, Environment, Ecosystems And Sustainable Development Vouliagmeni, Athens, Greece, July 12-14, Available: http://www.wseas.us/e-library/conferences/2005athens/ee/papers/507-198.pdf

[21] E. F. Loshani, M. Sharifkhanim *An Optimum Solution for Telemetry of Distributed Wells in South of Tehran,* Proceedings of the 8th WSEAS International Conference on Signal Processing, Robotics And Automation, Cambridge, UK, February 21-23, 2009

[22] E.G. Ularu, F. Puican, A. Apostu, G. Suciu, A. Vulpe, *Analytical databases for the Cloud and Virtualization,* Proceedings of the IE 2013 International Conference, Bucharest, April 2013.

[23] Parts of the current paper have been published in Conference Proceedings: A. Apostu, F. Puican, G. Ularu, G. Suciu, G. Todoran,

Study on advantages and disadvantages of Cloud Computing – the advantages of Telemetry Applications in the Cloud, 13th WSEAS International Conference on Applied Computer Science(Iwate - ACS '13) - ISI Proceedings, 23-25 April 2013, Morioka City, Iwate, Japan.

Anca APOSTU has graduated The Academy of Economic Studies from Bucharest (Romania), Faculty of Cybernetics, Statistics and Economic Informatics in 2006. She has a Master diploma in Economic Informatics from 2010 and in present she is a PhD. Candidate in Economic Informatics with the Doctor's Degree Thesis: "Informatics solution in a distributed environment regarding unitary tracking of prices". Her scientific fields of interest include: Economics, Databases, Programming, Information Systems, Information Security and Distributed Systems.

Florina Camelia PUICAN is a PhD. Student, in the third year, at the Institute of Doctoral Studies. Bucharest. In 2008, she graduated from Faculty of Business Administration with teaching in foreign languages (English), at the Academy of Economic Studies, Bucharest and in 2009, from Faculty of Mathematics and Computer Science, section Computer Science, University of Bucharest. From 2010, she holds a Master Degree obtained at Faculty of Business Administration with teaching in foreign language (English), at the Academy of Economic Studies, Bucharest. During her studies and work experience she undertook a wide range of skills in economics, information technology and information systems for business, design and management of information systems and databases.

Elena-Geanina ULARU graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2008. She holds a Master Degree obtained at Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies at the Academy of Economic Studies and is currently a Phd. Student, in the second year, at the Institute of Doctoral Studies, doing her research at the University of Economics from Prague.

George SUCIU graduated from the Faculty of Electronics, Telecommunications and Information Technology at the University "Politehnica" of Bucharest in 2004. He holds a Master diploma in Informatics Project Management from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies Bucharest from 2010 and currently, his PhD work is focused on the field of Electronics Engineering and Cloud Communications. Also he is IEEE member and has received a type D IPMA certification in project management from Romanian Project Management Association / IPMA partner organization. He is the author or co-author of over 30 journal articles and scientific papers at conferences. His scientific fields of interest include: project management, electronics and telecommunication, cloud computing, big data, open source, IT security, data acquisition and signal processing. Known languages: German, French, English; He also has experience in project leading and participation in various research projects (FP7, National Structural Funds), with more than 15 years activity in information and telecommunication systems.

Gyorgy TODORAN has graduated the Faculty of Electronics, Telecommunications at University “Politehnica” in Bucharest in 2000. He holds a Master degree in Quality Management (2001) and Strategic Management (2002) from the “Politehnica” University of Bucharest. Currently he is working on his Ph.D. thesis in security technologies with focus on open source, cloud computing, mobile and BYOD initiatives. He has more than 10 years of experience in commercial and governmental telecommunication systems, mainly in system administration, system management, design, project management, consulting.

Measuring Data Quality in Analytical Projects

Anca Ioana ANDREESCU, Anda BELCIU, Alexandra FLOREA,
Vlad DIACONITA
University of Economic Studies, Bucharest, Romania
anca.andreescu@ie.ase.ro, anda.velicanu@ie.ase.ro, alexandra.florea@ie.ase.ro,
diaconita.vlad@ie.ase.ro

Measuring and assuring data quality in analytical projects are considered very important issues and overseeing their benefits may cause serious consequences for the efficiency of organizations. Data profiling and data cleaning are two essential activities in a data quality process, along with data integration, enrichment and monitoring. Data warehouses require and provide extensive support for data cleaning. These loads and renew continuously huge amounts of data from a variety of sources, so the probability that some of the sources contain "dirty data" is great. Also, analytics tools offer, to some extent, facilities for assessing and assuring data quality as a built in support or by using their proprietary programming languages. This paper emphasizes the scope and relevance of a data quality measurement in analytical projects by the means of two intensively used tools such as Oracle Warehouse Builder and SAS 9.3.

Keywords: *data quality, data profiling, analytical tools, data warehouses*

1 Introduction

Data quality represents an important issue in every business. To be successful, companies need high-quality data on inventory, supplies, customers, vendors and other vital enterprise information in order to run efficiently their data analysis applications (e.g. decision support systems, data mining, customer relationship management) and produce accurate results. As companies develop analytical and business intelligence systems on their transactional systems, the reliability of key performance indicators and data mining predictions will depend entirely on the validity of the data on which they are based. Any type of data quality issue could potentially lead to erroneous data mining and analysis results which in turn could lead to severe consequences, financial or otherwise.

But while the importance of valid data for business decision making is increasing, so does to the same extent the challenge to ensure their validity. Information flows continuously in the company from various sources and systems and a large number of users and so the volume of data being generated is increasing exponentially day by day.

2. Data Quality Assessment

Data quality assessment is a complex process through which we can obtain a complete assessment of an organization’s data. Through this process we get not only a full image regarding the data quality issues the company is facing but also an accurate view of the time and effort required to fix those problems.

In the first stages of research on data quality an evaluation method using a vector composed of elements which describe several easily evaluated data quality aspects was proposed [1]. Amongst the quality factors included we mention relevance, accuracy, data actuality etc. Later on, the list of proposed quality factors by researchers such as [2], [3] and [4] grew larger and larger reaching almost 200 elements.

However, defining a number of sophisticated data quality factors, how large that number might be, is not enough to obtain a relevant data quality assessment. As mentioned in [5] data quality assessment is highly context and application dependent. It is difficult to formulate a general solution that will work in all situations.

The metrics for determining data quality may be difficult to define because they are domain or application specific. A common method to define the quality of the data is represented by data profiling.

Data profiling is the process of analyzing large data sets obtaining a set of statistical indicators regarding that data, such as minimum, maximum, mean, mode, percentile, standard deviation, frequency, and variation as well as other aggregates such as count and sum. During data profiling we could obtain additional metadata information such as data type, length, discrete values, uniqueness, occurrence of null values, typical string patterns, and abstract type recognition [6]. Through data profiling we make an assessment of data to understand its content, structure, quality and dependencies.

There are some common methods used in profiling, no matter the tool selected, as mentioned in [7]: structure discovery – verification if the different patterns of data are valid; data discovery – verifies if the data value is correct, error free and valid; relationship discovery – checking if all the key relationships are maintained and data redundancy where we check if the same data has multiple representations.

Profiling techniques can be grouped in two categories: manual or automated using a profiling tool. Manual techniques involve people who unravel the data to assess their condition, query by query. This is appropriate for small data sets from a single source, with less than 50 fields, where data is relatively simple.

The automated techniques use software tools to gather summary statistics and analysis. These tools are most appropriate for projects with hundreds of thousands of records, many fields, multiple sources and questionable documentation and metadata. Sophisticated technology has been built for data profiling to handle complex problems, particularly for high-profile projects and critical missions.

Choosing an appropriate profiling tool might be a difficult task so it’s useful to know what the differences between them are. Mostly they vary in the architecture used to analyze the data and in the work environment they provide for the team that generates the data profile.

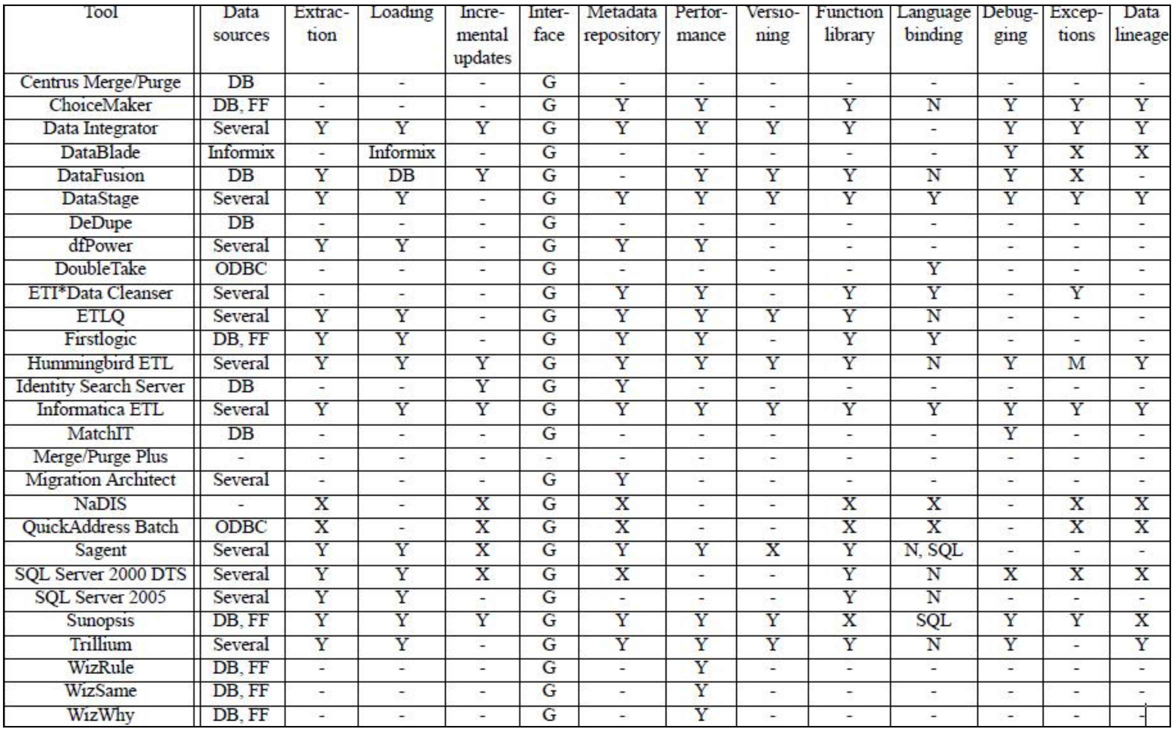
From the architectural point of view we distinguish between query based profiles and repository based profiles.

Some profiling tools require having technical skill at running SQL queries on source data or a view of the source data. Although this creates good information about the data it also has several limitations regarding performance:

- **Performance risks** – the queries operated on production systems slow the systems, sometimes significantly. When additional information is needed or if users want to see the actual data, a second query is executed creating more pressure on the system. This risk can be reduced by making a copy of the data, but this requires replicating the entire environment, both hardware systems and software, which can be costly and time consuming.
- **Traceability risks** - the data in production systems is constantly changing. Statistics and metadata extracted from profiles based on queries risk to immediately become outdated.
- **Integrity risks** – it is complicated to acquire comprehensive knowledge using query-based analysis. The queries are based on assumptions, and the objective is to confirm and quantify expectations about what is wrong and right in the data. Given this, it is easy to overlook problems that you have not reported.

Other tools generate data profiles as part of a scheduled process and store the results in a profile repository. Saved results can include content such as summary statistics, metadata, patterns, key relationships and data values. These results can be further analyzed by the users or be saved for later trend analysis. Profile repositories that allow users to explore information and view the values of the original data in the context of source records are those that offer more versatility and stability for non-technical audiences.

As mentioned before, there is a great number of data quality and profiling tools available on the market and choosing between them can be a difficult task. In [8] was conducted an extensive analysis of the major features of a series of data quality tools which is summarized in figure 1.



Tool	Data sources	Extrac-tion	Loading	Incre-mental updates	Inter-face	Metadata repository	Perfor-mance	Versio-ning	Function library	Language binding	Debug-ging	Excep-tions	Data lineage
Centrus Merge/Purge	DB	-	-	-	G	-	-	-	-	-	-	-	-
ChoiceMaker	DB, FF	-	-	-	G	Y	Y	-	Y	N	Y	Y	Y
Data Integrator	Several	Y	Y	Y	G	Y	Y	Y	Y	-	Y	Y	Y
DataBlade	Informix	-	Informix	-	G	-	-	-	-	-	Y	X	X
DataFusion	DB	Y	DB	Y	G	-	Y	Y	Y	N	Y	X	-
DataStage	Several	Y	Y	-	G	Y	Y	Y	Y	Y	Y	Y	Y
DeDupe	DB	-	-	-	G	-	-	-	-	-	-	-	-
dfPower	Several	Y	Y	-	G	Y	Y	-	-	-	-	-	-
DoubleTake	ODBC	-	-	-	G	-	-	-	-	Y	-	-	-
ETI*Data Cleanser	Several	-	-	-	G	Y	Y	-	Y	Y	-	Y	-
ETLQ	Several	Y	Y	-	G	Y	Y	Y	Y	N	-	-	-
Firstlogic	DB, FF	Y	Y	-	G	Y	Y	-	Y	Y	-	-	-
Hummingbird ETL	Several	Y	Y	Y	G	Y	Y	Y	Y	N	Y	M	Y
Identity Search Server	DB	-	-	Y	G	Y	-	-	-	-	-	-	-
Informatica ETL	Several	Y	Y	Y	G	Y	Y	Y	Y	Y	Y	Y	Y
MatchIT	DB	-	-	-	G	-	-	-	-	-	Y	-	-
Merge/Purge Plus	-	-	-	-	-	-	-	-	-	-	-	-	-
Migration Architect	Several	-	-	-	G	Y	-	-	-	-	-	-	-
NaDIS	-	X	-	X	G	X	-	-	X	X	-	X	X
QuickAddress Batch	ODBC	X	-	X	G	X	-	-	X	X	-	X	X
Sagent	Several	Y	Y	X	G	Y	Y	X	Y	N, SQL	-	-	-
SQL Server 2000 DTS	Several	Y	Y	X	G	X	-	-	Y	N	X	X	X
SQL Server 2005	Several	Y	Y	-	G	-	-	-	Y	N	-	-	-
Sunopsis	DB, FF	Y	Y	Y	G	Y	Y	Y	X	SQL	Y	Y	X
Trillium	Several	Y	Y	-	G	Y	Y	Y	Y	N	Y	-	Y
WizRule	DB, FF	-	-	-	G	-	Y	-	-	-	-	-	-
WizSame	DB, FF	-	-	-	G	-	Y	-	-	-	-	-	-
WizWhy	DB, FF	-	-	-	G	-	Y	-	-	-	-	-	-

Fig.1. General functionalities of commercial data quality tools [8]

The authors have used the following notations for constructing their feature analysis table: Y: supported; X: not supported; -: unknown information; N: native; DB: only relational databases; FF: only flat files; G: graphical; M: manual. Such an analysis can represent a major support element when choosing an appropriate data quality tool. Once the data profiling process is completed and all the problems have been identified, special attention must be paid to data cleaning. Through the cleansing (or scrubbing) of data we can detect and remove errors and inconsistencies and thus improve the quality of data.

The complexity of the data cleaning process varies due to the type of storage of the processed data. If data is stored in single collections the problems that might arise come from incorrect data entry, missing information or other invalid data. If data is to be integrated from multiple sources one of the main problems is redundancy: same data is stored in different representations. Thus it becomes necessary to eliminate duplicate information and consolidate different data representations in order to provide access to accurate and consistent data. When working with data warehouses, which process particularly large amounts of data from a array of sources extensive support for data cleaning is a mandatory requirement, especially since data warehouses are used for decision making, so that the correctness of their data is vital to avoid wrong conclusions. For instance, as mentioned in [9], duplicated or missing information will produce incorrect or misleading statistics ("garbage in, garbage out").

When selecting a data cleaning approach there are several requirement that should be taken into consideration: detection and removal of all major inconsistencies and errors in individual and multiple data sources; use of appropriate tools in order to limit manual inspection and programming effort; use of tools that can easily cover additional sources as they might appear; specification of mapping functions in a declarative and reusable way for other data sources as well as for query processing.

3. Data Profiling in Oracle Warehouse Builder

Oracle Warehouse Builder (OWB) has a Client - Server architecture, the database being stored on the Server and the Design Centre and the Repository Browser being available for the Client. The Core Features include Enterprise ETL (Extract Load Transform), Data Quality and ERP/CRM Connectors.

The way data Quality can be assured by OWB is shown in figure 2.

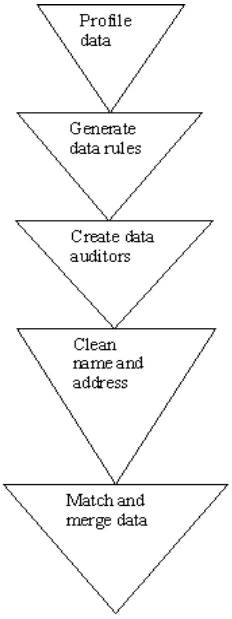


Fig. 2. Data Quality in QWB (adapted from source: [10])

Data profiling in OWB is "a systematic analysis of data sources" [10] in order to obtain new characteristics of data.

A data rule is an automatic or user - generated expression that allows data to be formatted according to domains, constraints, for it to gain its uniformity and consistency.

According to [10] "data auditors are process flows that evaluate one or more data rules against a given table".

Name and address cleansing assumes some transformations on these types of data in order to improve the quality. The transformations include parsing, standardization, augmentation, division, etc.

Matching and merging of data has the role of determining which values actually refer to the same logic data. This process helps eliminate duplicates and unite data in single row records.

The five steps in achieving data quality are profiling the data, generating data rules, deploying corrections and cleaning the data, as shown in figure 3.

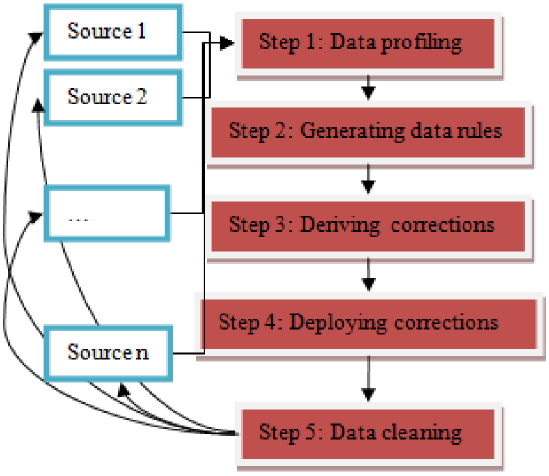

Source 1
Source 2
...
Source n
Step 1: Data profiling
Step 2: Generating data rules
Step 3: Deriving corrections
Step 4: Deploying corrections
Step 5: Data cleaning

Fig. 3. Steps in achieving data quality (adapted from source: [11])

The first and the last step work with data directly from its sources, but the cleaning process returns a new set of values.

Data Profiling is one of the ways for assuring Data Quality, next to Anomaly Detection, Business Rules and Audit, as Oracle sees it. Document [12] states the following: "Warehouse Builder enables to discover the structural content of data, capture its semantics, and identify any anomalies or outliers prior to loading it in a system. With data profiling, one can automatically derive business rules and mappings to clean data, derive quality indices such as Six Sigma, and use auditors to continuously monitor data quality." This way data profiling can be integrated in the ETL process buying time and using quality data.

The steps in performing data profiling are presented in [13] and include: creating data profile objects (which are metadata objects in Oracle Projects), creating data profiles, configuring data profiles, loading all types of configuration parameters (pattern, domain, unique key, functional dependency, redundant column etc.), profiling data.

The main types of data profiling are described in figure 4**.**

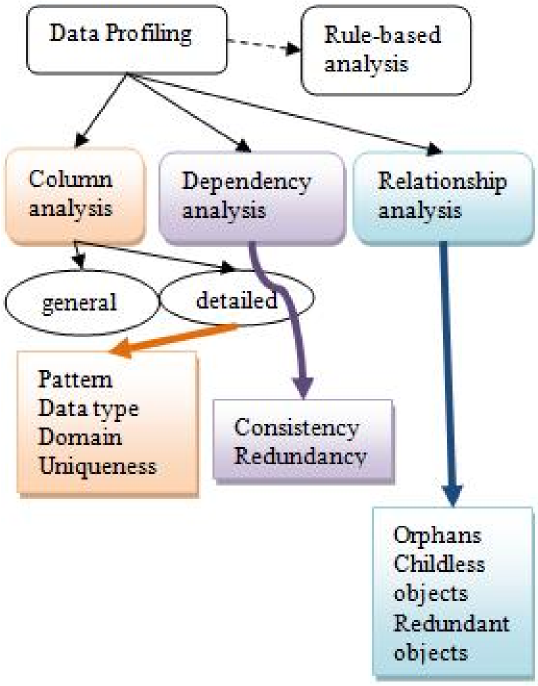


Fig.4. Types of data profiling (adapted from source: [14])

According to [11] data quality process should include four types of analyses as described below:

1. Column analysis that is based on: Uniqueness (metadata analysis); Completeness (the missing or incorrect values of attributes and thus the incomplete entities); Precision (precision and scale of numeric attributes); Uniformity (format analysis of numeric, character and date time attributes).
2. Dependency analysis, that consists of the following sub-analysis: Consistency of data type, length and domain; Primary key uniqueness; Redundancy avoided by using normalization.
3. Relationship analysis is based on: Referential integrity; Correctness using statistical control (minim, maxim, average, median etc.).
4. Rule-based analysis made through business and data rules.

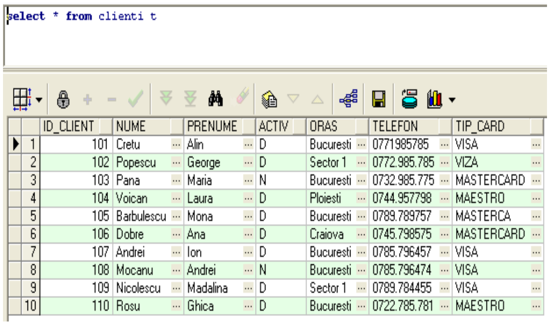
The advantages of using Oracle Warehouse Builder for data profiling are:

- allows discovering hidden things about apparently common data like: anomalies, additional relations between tables, patterns, complete domain of values, etc;
- the user can view the results in tabular or graphical format in Data Profile Editor;
- generates corrective ETL process based on business rules;
- has a simple and flexible design and generates robust ETL processes;
- works with a single repository that drives ETL processes and reporting.

The limitations of using OWB for data profiling are also shown below:

- the database workspace that is used must be Oracle 10g or higher and even though data could be accessed through JDBC connectivity, it must be staged first in an Oracle database and then used for profiling;
- data profiling process can analyze columns at a limit of 165 in each table. If it is necessary to analyze more at a time, an attribute set can be created in order to group more columns;
- complex data types cannot be analyzed if they are located on different database instances.

By following these analyses, we present an example of assuring data quality through data profiling made in Oracle Warehouse Builder. The initial set of data is presented in figure 5.



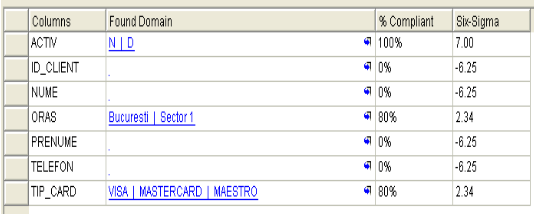
```
select * from clienti t
```

	ID_CLIENT	NUME	PRENUME	ACTIV	ORAS	TELEFON	TIP_CARD
1	101	Cretu	Alin	D	Bucuresti	0771985785	VISA
2	102	Popescu	George	D	Sector 1	0772.985.785	VIZA
3	103	Pana	Maria	N	Bucuresti	0732.985.775	MASTERCARD
4	104	Voican	Laura	D	Ploiesti	0744.957798	MAESTRO
5	105	Barbulescu	Mona	D	Bucuresti	0789.789757	MASTERCA
6	106	Dobre	Ana	D	Craiova	0745.798575	MASTERCARD
7	107	Andrei	Ion	D	Bucuresti	0785.796457	VISA
8	108	Mocanu	Andrei	N	Bucuresti	0785.796474	VISA
9	109	Nicolescu	Madalina	D	Sector 1	0789.784455	VISA
10	110	Rosu	Ghica	D	Bucuresti	0722.785.781	MAESTRO

Fig. 5. Initial set of data

First we import the metadata in Oracle Warehouse builder, after defining the connection. Once they have been imported we build a profile. Based on the imported data Oracle estimates data types and restrictions, which can be accepted and/or refined by the user.

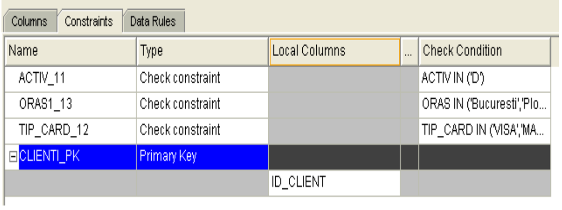
Domain values are also detected (figure 6). On this basis rules can be derived that will be later applied as data integrity restrictions. Oracle Data Warehouse can correct the data that do not meet these rules.



Columns	Found Domain	% Compliant	Six-Sigma
ACTIV	N | D	100%	7.00
ID_CLIENT	.	0%	-6.25
NUME	.	0%	-6.25
ORAS	Bucuresti | Sector 1	80%	2.34
PRENUME	.	0%	-6.25
TELEFON	.	0%	-6.25
TIP_CARD	VISA | MASTERCARD | MAESTRO	80%	2.34

Fig. 6. Data domains

We derive rules from these domains and accept only *Visa, Mastercard* and *Maestro* as card types. Also we accept *Bucureşti, Ploieşti, Craiova* as cities, the rest are mapped to unknown, the sectors of Bucharest will be corrected later. Also we say that only active clients are allowed in our analysis. Next we create the corrections. This are the corrections applied to the source data before being copied to the destinations. The bases for this are the derived data rules previously defined. As shown in figure 7 we can define different constraints (the check constraints are autodectected).



Columns | Constraints | Data Rules

Name	Type	Local Columns	...	Check Condition
ACTIV_11	Check constraint			ACTIV IN ('D')
ORAS1_13	Check constraint			ORAS IN ('Bucuresti','Plo...
TIP_CARD_12	Check constraint			TIP_CARD IN ('VISA','MA...
⊟CLIENTI_PK	Primary Key			
		ID_CLIENT		

Fig. 7. Define constraints

We next specify the action and the cleanse strategy for the corrections. We choose to remove the rows that aren't for active customers, correct the card type and the city. The similarity Match uses the built-in Match-Merge functionality in Oracle Warehouse Builder to change the erroneous value to the one that is most similar to it within the column domain. To correct the city we use a custom strategy and we add another function to correct the telephone number. The source code for these functions in the following:

```
//city correction
begin
if lower(oras) like ‘%sector%’
then return ‘Bucuresti’;
else return oras;
end if;

//telephone number correction
v_telefon varchar2(20);
BEGIN
For i in 1..lenght(telefon) loop
If substr(telefon,i,1) in
(‘0’,‘1’,
‘2’,‘3’,‘4’,‘5’,‘6’,‘7’,‘8’,‘9’)
then v_telefon:= v_telefon||
substr(telefon,i,1);
end if;
end loop;
RETURN v_telefon;
EXCEPTION
        WHEN OTHERS THAN NULL;
RETUN NULL:
END;
```

If we examine the mapping that implements the correction, you will note that the mapping first reads data from the original table, and then attempts to load it into a staging copy of the table with the data rule applied to it.

Those rows that pass the data rule are then copied into the corrected table. Those that fail any of the rules are then cleansed via pluggable mapping that allows you to take a series of mapping steps and "plug" them into another mapping (figure 8).

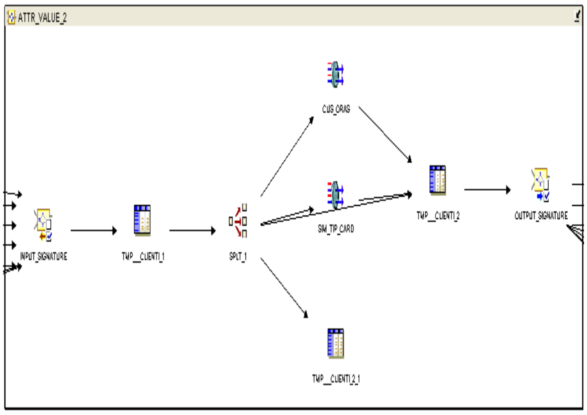
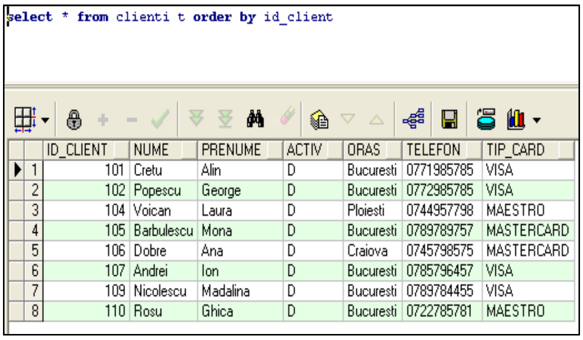


Fig. 8. Pluggable mapping

We deploy the correction objects, transformations, and mappings and then run the correction mapping.

The corrected data are to be found in the target schema as shown in figure 9**.**



```
select * from clienti t order by id_client
```

	ID_CLIENT	NUME	PRENUME	ACTIV	ORAS	TELEFON	TIP_CARD
1	101	Cretu	Alin	D	Bucuresti	0771985785	VISA
2	102	Popescu	George	D	Bucuresti	0772985785	VISA
3	104	Voican	Laura	D	Ploiesti	0744957798	MAESTRO
4	105	Barbulescu	Mona	D	Bucuresti	0789789757	MASTERCARD
5	106	Dobre	Ana	D	Craiova	0745798575	MASTERCARD
6	107	Andrei	Ion	D	Bucuresti	0785796457	VISA
7	109	Nicolescu	Madalina	D	Bucuresti	0789784455	VISA
8	110	Rosu	Ghica	D	Bucuresti	0722785781	MAESTRO

Fig. 9. The corrected data

We notice the profile of clients that have active accounts (two of them were eliminated), their city and phone number were corrected and the type of cards were adjusted to the new domain rules.

We can conclude that be using OWB one can easily define rules beginning by specifying the source and the target of the data. Domain values are detected providing a base for rules that can define the actions to take when data doesn't comply with the domain.

4. SAS Analytic Tools for Data Profiling

SAS has been known for many years as an important player in the market of the business analytics tools. It offers a variety of powerful software tools specialized in data management, data integration, analytics and reporting.

According to [15], across its solutions, SAS includes a large variety of analytical features, and therefore, it is not surprising that many of its functions are perfectly suited to profile and improve data quality. The extensive tool support for data quality in SAS can be broadly classified in two classes:

a) Built-in features offered by tools such as SAS Enterprise Miner, SAS Text Miner, SAS Model Manager, SAS Forecast Server, JMP and Data Flux Data Management Platform. These tools however, may not be available for some SAS users, may require additional training, and may be overkill if an understanding of the content of a file is all that is needed; that is, no data cleansing or other transformations are required [16].

b) Capabilities of SAS language packages, such as Base SAS, SAS/STAT and SAS/ETS.

By its analytics tools and functions, from SAS offers a variety of methods for data profiling that allow better insight into the data quality status and ways to improve it [15], such as:

- Outliners can be detected with statistical measures, while a most plausible value can be detected and calculated.
- Missing values can be imputed with methods varying from simple mean imputation to predictive models.
- Use of mathematical formulas and statistical measures to transform distribution into a more appropriate shape.

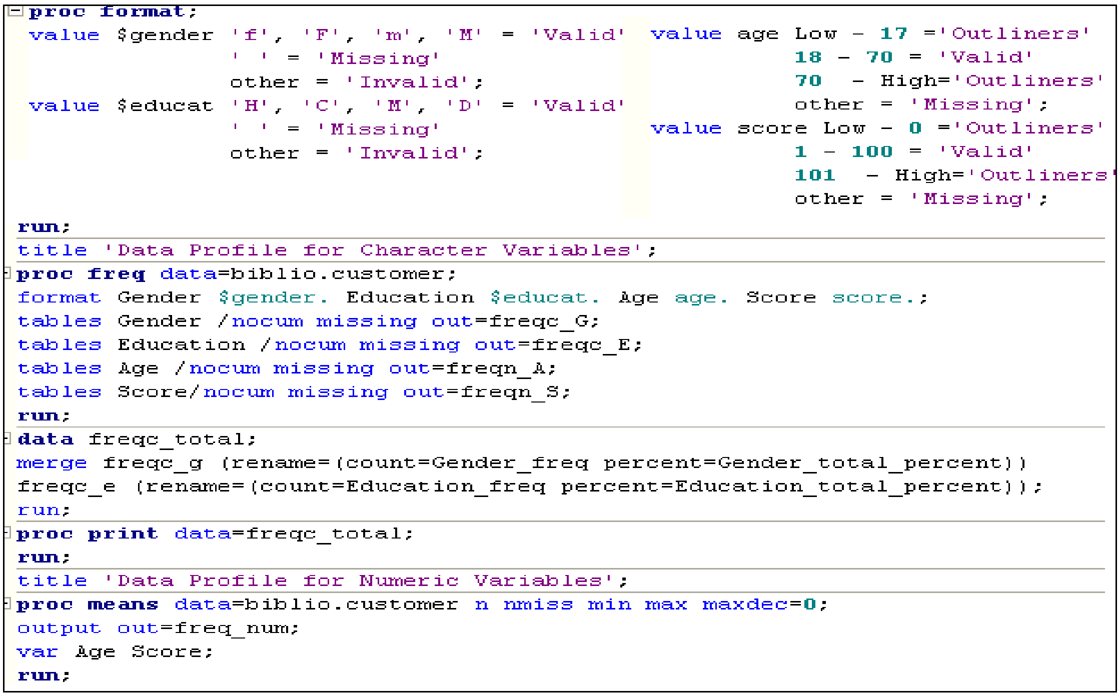
- Identify de-duplication of records based on analytical methods that describe the similarity and closeness of records.

In this paper we exemplify the use of Base SAS general quality control features that may be used to check data correctness and completeness. More precisely, we will make use of a SAS mechanism called format, which is a stored set of rules that can be used to restructure the cardinality of a column, either by viewing the data or by recoding the data [17]. These formats can be used to validate the data content through lookup tables for acceptable values for categorical variables or for acceptable ranges for interval variables.

It is worth mentioning that actually SAS formats have no exact analogy in other data management programming languages or analytic tools.

For demonstration purposes we consider a data set containing data collected from a customer satisfaction evaluation program. Four of the main variables included in the data set are Gender, Age, Education level and Score of evaluation. Valid values for these variables are described accordingly to the four SAS formats defined in the code below.

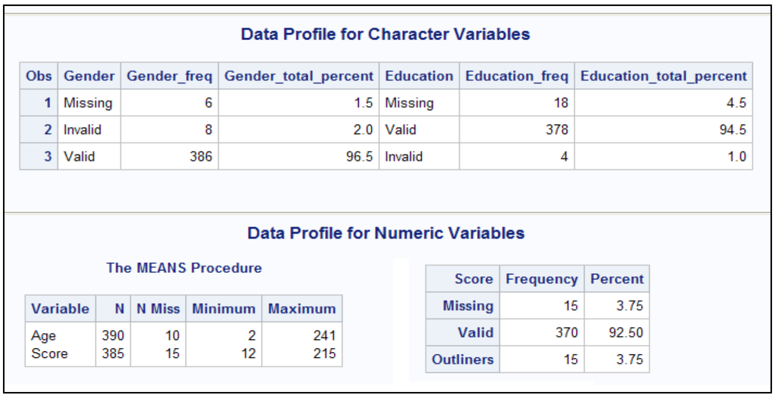
In combination with formats, PROC FREQ it is used in our example in order to determine number of the valid observations, missing values and invalid values both for character and numeric values. Afterwards, PROC MEANS it is used to obtain a data profile only for numeric variables. Below (figure 10) is the source code for creating a succinct data profile for these variables.



```
proc format;
 value $gender 'f', 'F', 'm', 'M'  = 'Valid'  value age Low - 17  ='Outliners'
               ' ' = 'Missing'                          18 - 70  = 'Valid'
               other = 'Invalid';                       70   - High='Outliners'
 value $educat 'H', 'C', 'M', 'D'  = 'Valid'            other = 'Missing';
               ' ' = 'Missing'                value score Low - 0  ='Outliners'
               other = 'Invalid';                       1 - 100  = 'Valid'
                                                        101   - High='Outliners'
                                                        other = 'Missing';
run;
title 'Data Profile for Character Variables';
proc freq data=biblio.customer;
format Gender $gender. Education $educat. Age age. Score score.;
tables Gender /nocum missing out=freqc_G;
tables Education /nocum missing out=freqc_E;
tables Age /nocum missing out=freqn_A;
tables Score/nocum missing out=freqn_S;
run;
data freqc_total;
merge freqc_g (rename=(count=Gender_freq percent=Gender_total_percent))
freqc_e (rename=(count=Education_freq percent=Education_total_percent));
run;
proc print data=freqc_total;
run;
title 'Data Profile for Numeric Variables';
proc means data=biblio.customer n nmiss min max maxdec=0;
output out=freq_num;
var Age Score;
run;
```

Fig.10. SAS source code for data profiling

Output for the above SAS code is presented in figure 11, indicating an overall data profile for both numeric and character variables. We can observe that the MEAN procedure, by returning the minimum and maximum values for numeric variables, helps in identifying existing outliners.



Data Profile for Character Variables

Obs	Gender	Gender_freq	Gender_total_percent	Education	Education_freq	Education_total_percent
1	Missing	6	1.5	Missing	18	4.5
2	Invalid	8	2.0	Valid	378	94.5
3	Valid	386	96.5	Invalid	4	1.0

Data Profile for Numeric Variables

The MEANS Procedure

Variable	N	N Miss	Minimum	Maximum
Age	390	10	2	241
Score	385	15	12	215

Score	Frequency	Percent
Missing	15	3.75
Valid	370	92.50
Outliners	15	3.75

Fig.11. SAS profiling output

5. Conclusions

Data quality assessment should always be taken into account when managing and analyzing data, especially when large data volumes are or heterogeneous data sources are involved. In this paper it has been pointed out that specialized tool for data warehouses and business analytics offer support for various stages of the data quality process, including data profiling and data validation. Different tools offer different kind of support in this regard, depending on their scope. From the two examples using Oracle and SAS software, we can conclude that there are three main approaches to data quality: 1) to use predefined tools facilities, which should be straightforward; 2) to use programming languages like PL/SQL or SAS to write specialized routines, which is more time consuming, but offers great flexibility; 3) to combine the above approaches in order to customize predefined tool support.

References

[1] Juliusz L. Kulikowski. “Data Quality Assessment”. *Handbook of Research on Innovations in database technologies and Applications: Current and Future Trends*, Information Science Reference publishing house, 2009, pp 378-384, ISBN13: 978-160-566-242-8.

[2] Leo L. Pipino, Yang W. Lee, Richard Y. Wang. “Data Quality Assessment”, Communications of the ACM - Supporting community and building social capital, Vol.45, Issue 4, pp. 211-218, April 2002.

[3] G. Shanks, P. Darke. “Understanding Metadata and Data Quality in a Data Warehouse”, Australian Computer Journal, Vol. 30, pp 122-128, 1998.

[4] Richard Y. Wang, Veda C. Storey, Christopher P. Firth. “A Framework for Analysis of Data Quality Research”, IEEE Transactions on Knowledge and Data Engineering, Vol. 7, No. 4, August 1995.

[5] Tamraparni Dasu. “Data Glitches: Monsters in your Data”. Handbook, Reference book, 2012. Available: http://www.research.att.com/techdocs/TD_100950.pdf

[6] David Loshin. “Master Data Management”, Morgan Kaufmann Publishers, pp. 94-96, ISBN 978-012-374-225-4.

[7] Sanjay Seth. “Data Quality Assessment Approach” Internet. Available: http://hosteddocs.ittoolbox.com/ss052809.pdf

[8] - José Barateiro, Helena Galhardas –„A survey of data quality tools”, Datenbank-Spektrum 01/2005; 14:15-21

[9] - Erhard Rahm, Hong Hai Do – „Data Cleaning: Problems and Current Approaches”, IEEE Data Engineering Bulletin, Volume 23, 2000

[10] “Oracle Warehouse Builder 10gR2 Transforming Data into Quality Information”, January 2006. Available: http://www.oracle.com/technetwork/developer-tools/warehouse/transforming-1.pdf

[11] E. Borowski, H.-J. Lenz, “Design of a workflow system to improve data quality using Oracle Warehouse Builder”, Journal

of Applied Quantitative Methods, vol. 3, no. 3, pp. 198-206, Fall 2008.

[12] “Oracle Warehouse Builder User's Guide 10g Release 2 (10.2.0.2)”, April 2009, http://docs.oracle.com/cd/B31080_01/doc/owb.102/b28223.pdf

[13] “Oracle Warehouse Builder Data Modeling, ETL, and Data Quality Guide 11g Release 2 (11.2)”. Available: http://docs.oracle.com/cd/E11882_01/owb.112/e10935/data_profiling.htm

[14] “Oracle Warehouse Builder User's Guide 11g Release 1 (11.1)”, July 2007. Available: http://isu.ifmo.ru/docs/doc111/owb.111/b31278.pdf

[15] G. Svolba, “Data Quality for Analytics Using SAS”. SAS Press, 2012, pp. 182-192.

[16] S. J. Nowlin. “Data Profiling Using Base SAS® Software: A Quick Approach to Understanding Your Data”. SUGI 31 Proceedings. San Francisco, California March 26-29, 2006. Available: http://www2.sas.com/proceedings/sugi31/161-31.pdf

[17] P. Welbrock. “Validating And Updating Your Data Using SAS Formats”. NESUG 2001, Baltimore, USA, 2001.

Anca Ioana ANDREESCU, PhD is an associate professor at the Bucharest University of Economic Studies, Faculty of Cybernetics, Statistics and Informatics, Department of Economic Informatics and Cybernetics. She published over 20 articles in journals and magazines in computer science, informatics and business management fields, over 30 papers presented at national and international conferences, symposiums and workshops. In January 2009 she finished the doctoral stage, the title of her PhD thesis being: The Development of Software Systems for Business Management. She is the author of one book and she is coauthor of five books. Her interest domains related to computer science are: requirements engineering, business analytics, modeling languages, business rules approaches and software development methodologies.

Anda BELCIU has graduated the Faculty of Economic Cybernetics, Statistics and Informatics of the Bucharest Academy of Economic Studies, in 2008. She has a PhD in Economic Informatics and since October 2012 she is a Lecturer. She teaches Database, Database Management Systems and Software Packages seminars and courses at the Economic Cybernetics, Statistics and Informatics Faculty. She is co-author of 4 books, has 11 articles published in prestigious journals included in international recognized databases (SCOPUS, Elsevier, EBSCO, ProQuest, or DOAJ) and also 17 papers in the volumes of national and international scientific manifestations, of which 4 are indexed Thomson ISI Web of Science. Her scientific fields of interest and expertise include database systems, e-business, e-learning, spatial databases. She has experience in 5 research projects, participating as a team member.

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

Vlad DIACONITA is a member of the IEEE and INFOREC organizations and member of the technical team of the Database Systems Journal. As part of the research team he has worked in 8 different phases of 3 UEFISCDI funded grants. He has published more than 30 papers in peer reviewed journals and conference proceedings, many indexed in ISI or SCOPUS. He is the co-author of four books.

Model-Based Testing: The New Revolution in Software Testing

Hitesh KUMAR SHARMA, Sanjeev KUMAR SINGH, Prashant AHLAWAT
[1]University of Petroleum and Energy Studies
[2]Galgotia University Noida
[3]GITM Gurgaon
hkshitesh@gmail.com, sksingh8@gmail.com, prashantahlawat@ymail.com

The efforts spent on testing are enormous due to the continuing quest for better software quality, and the ever growing complexity of software systems. The situation is aggravated by the fact that the complexity of testing tends to grow faster than the complexity of the systems being tested, in the worst case even exponentially. Whereas development and construction methods for software allow the building of ever larger and more complex systems, there is a real danger that testing methods cannot keep pace with construction, hence these new systems cannot be sufficiently fast and thoroughly be tested. This may seriously hamper the development of future generations of software systems.

One of the new technologies to meet the challenges imposed on software testing is model-based testing. Models can be utilized in many ways throughout the product life-cycle, including: improved quality of specifications, code generation, reliability analysis, and test generation.

This paper will focus on the testing benefits from MBT methods and review some of the historical challenges that prevented model based testing and we also try to present the solutions that can overcome these challenges.

Keywords: *MBT, Test Cases, SUT, Test Suite.*

1 Introduction

“Model-based testing is a testing technique where the runtime behavior of an implementation under test is checked against predictions made by a formal specification, or model.”[7]. The IEEE definition of testing is "the process of exercising or evaluating a system or system component by manual or automated means to verify that it satisfies specified requirements or to identify differences between expected and actual results." [8]. Software testing is the process of executing a software system to determine whether it matches its specification and executes in its intended environment. A software failure occurs when a piece of software does not perform as required and expected. In testing, the software is executed with input data, or test cases, and the output data is observed. As the complexity and size of software grows, the time and effort required to do sufficient testing grows. Manual testing is time consuming, labor-intensive and error prone. Therefore it is pressing to automate the testing effort. The testing effort can be divided into three parts: test case generation, test execution, and test evaluation.

However, the problem that has received the highest attention is test-case selection. A test case is the triplet [S, I, O] where I is the data input to the system, S is the state of the system at which the data is input, and O is the expected output of the system. The output data produced by the execution of the software with a particular test case provides a specification of the actual program behavior. Test case generation in practice is still performed manually most of the time, since automatic test case generation approaches require formal or semi-formal specification to select test case to detect faults in the code implementation. Code based testing not an entirely satisfactory approach to generate guarantee acceptably thorough testing of modern software products. Source code is no longer the single source for selecting test cases, and nowadays, we can apply testing techniques all along the development process, by basing test selection on different pre-code artifacts, such as requirements, specifications and design models [9],[10]. Such a model may be generated from a formal specification or may be designed by software engineers through diagrammatic tools. Code based testing has two important disadvantages. First, certain aspects of behavior of a system are difficult to extract from code but are easily obtained from design models. The state based behavior captured in a state diagram and message paths are simple examples of this. It is very difficult to extract the state model of a class from its code. On the other hand, it is usually explicitly available in the design model. Similarly, all different sequences in which messages may be interchanged among classes during the use of a software is very difficult to extract from the code, but is explicitly available in the UML sequence diagrams. Another prominent disadvantage of code based testing is very difficult to automate and code based testing overwhelmingly depends on manual test case design.

2. Process and Terminology

We use this section to fix terminology and describe the general process of model-based testing. A test suite is a finite set of test cases. A test case is a finite structure of input and expected output: a pair of input and output in the case of deterministic transformative systems, a sequence of input and output in the case of deterministic reactive systems, and a tree or a graph in the case of non-deterministic reactive systems. The input part of a test case is called test input. In general, test cases will also include additional information such as descriptions of execution conditions or applicable configurations, but we ignore these issues here.

3. Model Based Testing

A generic process of model-based testing then proceeds as follows (Fig. 1).

Step 1.

A model of the SUT is built on the grounds of requirements or existing specification documents. This model encodes the intended behavior, and it can reside at various levels of abstraction.

The most abstract variant maps each possible input to the output "no exception" or "no crash". It can also be abstract in that it neglects certain functionality, or disregards certain quality-of-service attributes such as timing or security.

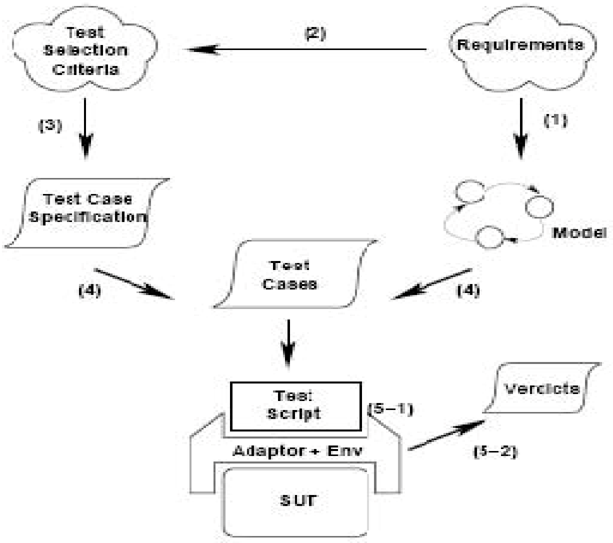


Fig. 1. The Process of Model-Based Testing

Step 2.
Test selection criteria are defined. In general, it is difficult to define a "good test case" a-priori. Arguably, a good test case is one that is likely to detect severe and likely failures at an acceptable cost, and that is helpful with identifying the underlying fault. Unfortunately, this definition is not constructive. Test selection criteria try to approximate this notion by choosing a subset of behaviors of the model. A test selection criterion possibly informally describes a test suite. In general, test selection criteria can relate to a given functionality of the system (requirements based test selection criteria), to the structure of the model (state coverage, transition coverage, def-use coverage), to stochastic characterizations such as pure randomness or user profiles, and they can also relate to a well-defined set of faults.

Step 3.
Test selection criteria are then transformed into test case specifications. Test case specifications formalize the notion of test selection criteria and render them operational: given a model and a test case specification, some automatic test case generator must be capable of deriving a test suite (see step 4). For instance, "state coverage" would translate into statements of the form "reach _" for all states _ of the (finite) state space, plus possibly further constraints on the length and number of the test cases. Each of these statements is one test case specification. The difference between a test case specification and a test suite is that the former is intensional ("fruit") while the latter is extensional ("apples, oranges, ..."): all tests are explicitly enumerated.

Step 4.
Once the model and the test case specification are defined, a test suite is generated. The set of test cases that satisfy a test case specification can be empty. Usually, however, there are many test cases that satisfy it. Test case generators then tend to pick some at random.

Step 5.
Once the test suite has been generated, the test cases are run (sometimes, in particular in the context of non-deterministic systems, generating and running tests are dove-tailed).
Running a test case includes two stages.

Step 5-1.
Recall that model and SUT reside at different levels of abstraction, and that these different levels must be bridged [2]. Executing a test case then denotes the activity of applying the concretized input part of a test case to the SUT and recording the SUT's output. Concretization of the input part of a test case is performed by a component called the adaptor. The adaptor also takes care of abstracting the output (see Fig 1).

Step 5-2.
A verdict is the result of the comparison of the output of the SUT with the expected output as provided by the test case. To this end, the output of the SUT must have been abstracted. Consider the example of testing a chip card that can compute digital signatures [7]. The verdict can take the outcomes pass, fail, and inconclusive. A test passes if expected and actual output conforms. It fails if they do not, and it is inconclusive when this decision cannot be made.

4. Importance of MBT

The first obstacle to overcome in developing tests is to determine the test target. While this may sound trivial, it is often the first place things go wrong. A description of the product or application to be tested is essential. The form the description can come in may vary from a set of call flow graphs for a voice mail system, to the user guide for a billing system's GUI. A defined set of features and / or behaviors of a product is needed in order to define the scope of the work (both development and test). The traditional means of specifying the correct system behavior is with English prose in the form of a Requirement Specification or Functional Specification [1]. The specification, when in prose, is often incomplete - only the typical or ideal use of the feature(s) is defined, not all of the possible actions or use scenarios. This incomplete description forces the test engineer to wait until the system is delivered so that the entire context of the feature is known. When the complete context is understood, tests can be developed that will verify all of the possible remaining scenarios. Another problem with textual descriptions is that they are ambiguous, (for example “if an invalid digit is entered, it shall be handled appropriately.”) The ‘appropriate’ action is never defined; rather, it is left to the reader’s interpretation.

5. Industry importance

Modeling is a very economical means of capturing knowledge about a system and then reusing this knowledge as the system grows. For a testing team, this information is gold; what percentage of a test engineer's task is spent trying to understand what the System Under Test (SUT) should be doing? (Not just is doing.) Once this information is understood, how is it preserved for the next engineer, the next release, or change order? If you are lucky it is in the test plan, but more typically buried in a test script or just lost, waiting to be rediscovered. By constructing a model of a system that defines the systems desired behavior for specified inputs to it, a team now has a mechanism for a structured analysis of the system. Scenarios are described as a sequence of actions to the system, with the correct responses of the system also being specified. Test coverage is understood and test plans are developed in the context of the SUT, the resources available and the coverage that can be delivered. The largest benefit is in reuse; all of this work is not lost. The next test cycle can start where this one left off. If the product has new features, they can be incrementally added to the model; if the quality must be improved, the model can be improved and the tests expanded; if there are new people on the team, they can quickly come up to speed by reviewing the model.

The increased complexity of systems as well as short product release schedules makes the task of testing challenging. One of the key problems is that testing typically comes late in the project release cycle, and traditional testing is performed manually. When bugs are detected, the cost of rework and additional regression testing is costly and further impacts the product release. The increased complexity of today’s software-intensive systems means that there are a potentially indefinite number of combinations of inputs and events that result in distinct system outputs and many of these combinations are often not covered by manual testing. We work with companies that have high process maturity levels, and excellent measurement data that shows that testing is more 50-75% of the total cost of a product release, yet these mature processes are not addressing this costly issue.

Test tools may not replace human intelligence in testing, but without them testing complex systems at a reasonable cost will never be possible. There are commercial products to support automated testing, most based on capture/playback mechanisms, and organizations that have tried these tools quickly realize that these approaches are still manually intensive and difficult to maintain. Even small changes to the application functionality or GUI can render a captured test session useless. But more importantly, these tools don't help test organizations figure out what tests to write, nor do they give any information about test coverage of the functionality.

6. Challenges

The real work that remains for the foreseeable future is fitting specific models (finite state machines, grammars or language-based models) to specific application domains. Often this will require new invention as mental models are transformed into actual models. Perhaps, special purpose models will be made to satisfy very specific testing requirements and more general models will be composed from any number of pre-built special-purpose models**.**

- Finding suitable abstractions is difficult
- We cannot execute partial tests

7. How can we overcome from these challenges

Fortunately, many of these problems can be resolved one way or the other with some basic skill and organization. Alternative styles of testing need to be considered where insurmountable problems that prevent productivity are encountered. We must form an understanding of how we are testing and be able to sufficiently communicate that understanding so that testing insight can be encapsulated as a model for any and all to benefit from. To achieve these goals, models must evolve from mental understanding to artifacts formatted to achieve readability and reusability. We must form an understanding of how we are testing and be able to sufficiently communicate that understanding so that testing insight can be encapsulated as a model for any and all to benefit from.

8. Conclusion

There is promising future for MBT as software becomes even more ubiquitous and quality becomes the only distinguishing factor between brands. When all vendors have the same features, the same ship schedules and the same interoperability, the only reason to buy one product over another is quality. MBT, of course, cannot and will not guarantee or even assure quality. However, its very natural, thinking through uses and test scenarios in advance while still allowing for the addition of new insights, makes it a natural choice for testers concerned about completeness, effectiveness and efficiency.

References

[1] IEEE standard for Requirements Specification (IEEE/ANSI Std. 830-1984) , IEEE Computer Society, (830-1993) IEEE Recommended Practice for Software Requirements Specifications (ANSI), IEEE Standard for Software Unit Testing (ANSI), IEEE Standard for Software Verification and Validation Plans (ANSI) found at: http://standards.ieee.org/catalog/it.html

[2] Proceedings of the Third IEEE International Symposium on Requirements Engineering, IEEE Computer Society, 1997.

[3] Paulk, M., Curtis, B., Chrissis, M.B., and Weber, C., Capability Maturity Model, Version 1.1 The Software Engineering Institute, Carnegie Mellon University. Found at: http://www.sei.cmu.edu/products/publications/96.reports/96.ar.cmm.v1.1.html

[4] ITU-T. ITU -T Recommendation Z.100: Specification and Description Language (SDL). ITU-T, Geneva, 1988. More can be found at http://www.sdl-forum.org/

[5] Spivey, M., The Z Notation: A Reference Manual, Second Edition. Prentice-Hall International, 1992.

[6] Beizer, B., *Black Box Testing*, New York, John Wiley & Sons, 1995. ISBN 0-471-12094-4.

[7] A. Pretschner, W. Prenninger, S. Wagner, C. Kuhnel, M. Baumgartner, B. Sostawa, R. Z¨olch, T. Stauner, One evaluation of model based testing and its automation, in: Proc. ICSE'05, 2005, pp. 392–401.

[8] R. Helm, I. M. Holland, and D.Gangopadhyay. Contracts: specifying behavioral compositions in object-oriented systems. In

Proceedings of the 5th Annual Conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA ’90), ACM SIGPLAN Notices, 25(10):169–180, 1990.

[9] A. Pretschner, J. Philipps, Methodological Issues in Model-Based Testing, in: [29], 2005, pp. 281–291.

[10] J. Philipps, A. Pretschner, O. Slotosch, E.Aiglstorfer, S. Kriebel, K. Scholl, Model based test case generation for smart cards, in: Proc. 8th Intl. Workshop on Formal Meth. For Industrial Critical Syst., 2003, pp. 168–192.

Dr. Hitesh Kumar Sharma, The author is an Assistant Professor (Senior Scale) in University of Petroleum & Energy Studies, Dehradun. He has published 20+ research paper in International Journal and 10+ research papers in National Journals. He is Ph.D. in Computer Science & Engineering.

Dr. Sanjeev Kumar Singh: The author is an Associate Professor in Galgotias University, Noida. He has published 30+ research paper in International Journal and 15+ research papers in National Journals. He is Ph.D. in Mathematics.

Mr. Prashant Ahlawat: The author is an Assistant Professor in GITM Gurgaon. He has published 10+ research paper in International Journal and 5+ research papers in National Journals. Currently He is pursuing his Ph.D. in Computer Science & Engineering.

Big Data: present and future

Mircea Răducu TRIFU, Mihaela Laura IVAN
University of Economic Studies, Bucharest, Romania
trifumircearadu@yahoo.com, ivanmihaela88@gmail.com

The paper explains the importance of the Big Data concept, a concept that even now, after years of development, is for the most companies just a cool keyword. The paper also describes the level of the actual big data development and the things it can do, and also the things that can be done in the near future.
The paper focuses on explaining to nontechnical and non-database related technical specialists what basically is big data, presents the three most important V's, as well as the new ones, the most important solutions used by companies like Google or Amazon, as well as some interesting perceptions based on this subject
Keywords: *Big Data, domains, risks, resources, information*

1 Introduction

Today world is totally and continuously connected to at least a big data informatics system, from support services, social network or a service that provide GPS localization. Every time you connect the phone to internet, or even pay with a card to get a soda you leave behind yourself traces full of information, using those traces you leave behind, any marketing department can and will know what is your destination and what are your habits.

Think yourself like a dear in the woods, you have many need and you what to satisfy them, you do so using at least one informatics tool, so when you search for the perfect bar it is like you shout in the woods and explore for the near water source. And like in any woods, there are some kind of friends, or a smells or sounds that can help you to find the best water source, in the real word that friend is can be any map provider, and the smell of the water is represented by any commercial you see.

The Big Data concept represents a in essence an "ocean of data" [5], lots of information and the means to analyses them.

In the present days most of the humans can access more information than most of our ancestors in a lifetime. Nowadays we double, in every year, the amount of data that we create. According to the global market intelligence firm IDC [5], in 2011 we played, swam, wallowed, and drowned in 1.8 zeta bytes of data.

Big data has many characteristics that made this term unique, big data is a concept that integrates all kinds of data, not just some basic ones like in a normal data warehouse, from text to pictures, sounds, movies, music, satellite coordinates and basically all kinds of input or output data that came from different types of sensors

Comparing with the actual flow of data, even Alexandria great library could be held at most 70.000 scrolls.

There are many ways to describe the concept. You can define it like the ability to extract meaning from this "ocean of data": to sort through masses of data and to find the hidden patterns and unexpected correlations.

The key of big data is not just to overflow servers with data, but use different types of algorithms that can use text and graphical

2. Defining Big Data

Forbes defines big data like this:

Table 1. Business Intelligence VS BigData

Traditional BI	**Reporting Big Data Analysis**
Reporting tool like Cognos, SAS, SSIS, SSAS	Visualization tool like QlikView or Tableau
Sample data or specific historical data	Huge volume of data
Data from the enterprise	Data from external sources like social media apart from enterprise data
Based on statistics	Based on statistics and social sentiment analysis or other data sources
Data warehouse and data mart	OLTP, real-time as well as offline data
Sequential Computing	Parallel Computing using multiple machines
Query languages SQL, TSQL	Scripting Languages Java script, Python, Ruby and SQL
Specific type of data : txt, xml, xls, etc	Multiple kinds of data : pictures, sounds, text, map coordinates

"Big data is a collection of data from traditional and digital sources inside and outside your company that represents a source of ongoing discovery and analysis" [5] (Fig. 1)

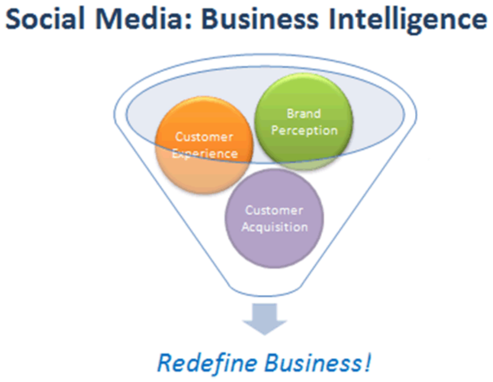


Fig. 1. Big data usage in BI

This is a more business approach, more practical and more business orientated than other definitions.

Big data is a mix of unstructured and multi structured data, those types of data are analyzed together to get more knowledge and information to company than could be get using the usual methods and infrastructure.

Unstructured data is information that is not organized or easily interpreted by traditional *data models* or *databases*, and usually is text-heavy. Good examples are posts from twitter, LinkedIn and other social media services. [5]

Multi-structured data is represented by a variety of data formats that came from interaction between peoples and machines, such as web applications and social services. Those include text and multimedia formats, like photos and videos, with structured data. [5]

Maybe the most use approach in defining Big Data is the one that was made by Gartner in 2001.According to Gartner Big data *is high volume*, *high velocity*, and/or high *variety* information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. This approach will be discussed in the next chapter [2].

3. Big data dimensions

The three V's of Gartners definition are: Volume, velocity and variety [9].

Volume: big data is that "Ocean of data" that we talk about in the rows above. It Is represented bay information that can came from every possible sensor, and some even say that we people are also sensors and data gatherers for *big data* [9].

The challenges of having such a big quantity of data is that is very hard to sustain it, to store it, to analyze it and ultimately to use it.

Velocity: is all about the speed of data traveling from one point to another and the speed of processing it. Sometimes it is crucial for the manager to be able to decide in a very little time on a variety of issues [9].

The most important issue is that the resources that analyses data is limited compared to the *volume* of data, but the requests of information is unlimited and usually information gets through at least one bottleneck.

Variety, the third characteristic is represented by the types of data that are stored. Because there are many types of sensors and sources, the data that came from them is varying very much in size and type. It is very complicated to analyze text, images and sounds in the same context and get a result that can be relied on.

And then is the issue of dark data, data that sits in the organization and is unused and also is not free.

There are one new dimension that were added to the existing ones: *Veracity* (Fig. 2)

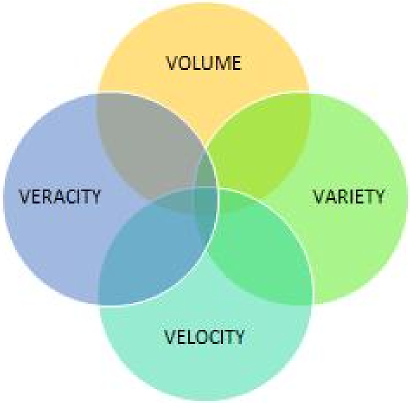


Fig. 2. Big data four V's

Veracity is the hardest thing to achieve with big data, because due to the *Volume* of information and the *variety* of its type is hard to identify the useful and accurate data from the "dirty data". The biggest problem is that the "dirty data" can lead very easy to an avalanche of errors, incorrect results and can affect the *Velocity* attribute of Big Data. The main purpose of the Big Data can be corrupted and all the information can lead to a useless and very expensive Big Data environment if there is not a good cleaning team.

The *Veracity* attribute is in its self also an objective for the Big Data developers. If the data cannot be accurate, is redundant or is unreliable, the whole Company can have a big problem, especial the companies that use big data to sell information like the marketing ones, or the ones that make market studies.

A lot of social media responses to campaigns could be coming from a small number of disgruntled past employees or persons employed by competition to post negative comments.

4. Big Data infrastructure

For a medium size or even big company that is not necessary making a living from renting space and processing power to clients, the construction of a Big Data infrastructure is often as expensive as inefficient.

So Big Data is not the answer for every type o company, is very expensive and hard do make on your own, and you need a specialized human resources. In the current labor market the Big Data specialists are very few and also the means to train programmers, architects and business analysts are few and very expensive.

The architecture of a Big Data solution is rather different from other data storage solution like Data Warehouse. The difference

It is basically represented by the four V's that characterizes the concept of big data.

One of the most important roles in the Big Data infrastructure is the NoSql Databases.

NoSql Databases

The term it means "Not Only SQL" rather than "No SQL", and it represents in essence a different kind of database approach, where the databases are not build with the relational databases structure, but use wide column store, document, key value structures or other types of structure that often are more easy to manage, and customize.

MongoDB is a document orientated, based on JSON, database that can handle large number of data sets with a low maintenance and that is easy to work with. [4]

Cassandra was originally a Facebook project, and after it was release as open source. It's one of the most important solutions and it has a huge community support. [4]

Cassandra is key and column orientated and is in many ways similar to the classic databases. It is also very close to the Google's Big Table, offering column indexes, strong support for denormalization and materialized views. [10]

BigTable is the solution used by Google, it is defined like a distributed store system used for managing structured data that is designed to a very large scale. [8]

BigTable achieves several goals some of them are wide applicability, scalability, high performance and high availability, and it is used today in more than sixty Google projects, like Google Analytics, Google Finance, Orkut, Personalized Search and Google Earth [8].

As a data model, Bigtable uses a sparse, distributed and persistent multidimensional sorted map. This map is indexed by row key, column key and timestamp so that every value in the map is an uninterpreted array of bytes. [8]

HBase is designed as an open soured clone to the BigTable, and is very similar in most of its models and designs, supports the same data structures tables.

HBase is integrated in the Hadoop project, so is easy to work using the database from a Map Reduce job.

MapReduce model

It is a programming model that has the purpose to process large data sets in parallel. [11]

The MapReduce model is using a pipeline that reads and writes to arbitrary file formats, with intermediate results been passed between stages as files, using computational spread across many machines, unlike the relational tables where all processing happens after the information has been loaded into a store, using specialized query language.[11]

Hadoop

It is a MapReduce system developed by Yahoo after the Google's MapReduce infrastructure (Fig. 3).

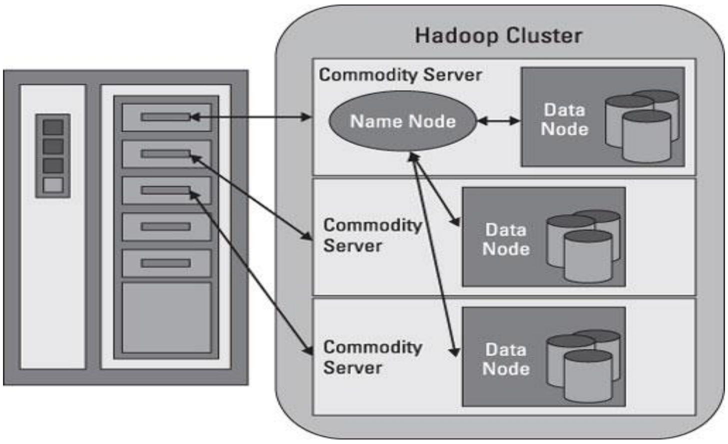


Fig.3. Hadoop Cluster [2]

To fully understand the capabilities of Hadoop MapReduce, we need to differentiate between MapReduce (the algorithm) and an implementation of MapReduce. Hadoop MapReduce is an implementation of the algorithm developed and maintained by the Apache Hadoop project. It is helpful to think about this implementation as a MapReduce engine, because that is exactly how it works. You provide input (fuel), the engine converts the input into output quickly and efficiently, and you get the answers you need.

Big data brings the big challenges of volume, velocity, and variety. As covered in the previous sections, HDFS resolves these challenges by breaking files into related collections of smaller blocks. These blocks are distributed among the data nodes in the HDFS cluster and then are managed by the Name Node. Block sizes are configurable and are usually 128 megabytes (MB) or 256MB, meaning that a 1GB file consumes eight 128MB blocks for its basic storage needs. HDFS is resilient, so these blocks are replicated throughout the cluster in case of a server failure. [2]

The HDFS keep track of all the pieces using metadata. The HDFS metadata provide detailed information like the following:

- Date of the creation, modification, execution of a file;
- Date where the blocks of the file are stored on the cluster;
- The rights to view or modify the file;
- Number many files are stored in one cluster;
- Number many data nodes exists in the cluster;
- Location address of the transaction log of the cluster.

JSON, we cannot continue the exemplification of some of the technologies that are used in the Big Data infrastructure without presenting one of the most popular formats for data processing.

Most of its popularity came from the easiness of reading and writing of both humans and machines. It is based on a JavaScript and is built on two structures: a collection of name/value pairs and an ordered list of values.

Big data infrastructure has multiple levels according to Michael Driscoll [3], who wrote about stack level of big data as can be seen in Fig. 4.

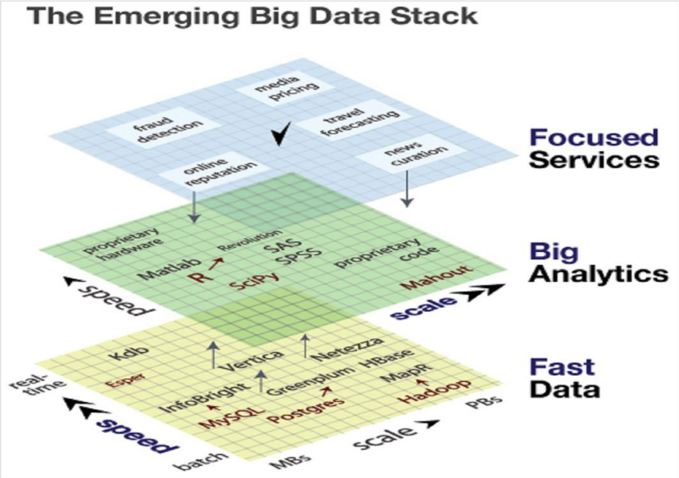


Fig. 4. Emerging big data stack [3]

Fast Data: At the base of the big data stack, where data is stored, processed, and queried the dominant axis of competition was once scale. But as cheaper commodity disks and Hadoop have effectively addressed scalable persistence and processing, the focus of competition has shifted toward speed. [3]

Big analytics: At the second tier of the big data stack, analytics is the brain to cloud computing. Here, however, the speed is less of a challenge; given an addressable data set in memory, most statistical algorithms can yield results in seconds. The challenge is scaling these out to address large datasets, and rewriting algorithms to operate in an online, distributed manner across many machines.

Focused services: The top of the big data stack is where data products and services directly touch consumers and businesses. For data start-ups, these offerings more frequently take the form of a service, offered as an API rather than a bundle of bits [3].

5. Uses of Big data

Scientific research has been revolutionized by Big Data. The Sloan Digital Sky Survey has today become a central resource for astronomers the world over. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database

Big Data has the potential to revolutionize not just research, but also education. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance.

Big Data in Medical industry

In the medical industry we can see big changes represented by the way data is now used not only to sell better medicines, or to make more profit but to increase the population access to hospitals and medical help as well. [5]

There are more and more cases where people like Yasmine Delwari Johnson, decides to take medical testes to find out what will be the characteristics of their child. They can now find out what are the odds that her child will have green eyes, or if he has big chances to develop lactose intolerance. [5]

Another use in the medical industry is to predict the medical condition of the patients based on their medical records, or their family.

Big Data Marketing

Is a process of collecting, analyzing, and executing on insights you've derived from big data to encourage customer engagement, improve marketing results, and measure internal account-ability.

Companies are focused on harnessing new data types and utilizing data to drive customer experience. Social media is driving most text analytics initiatives:

43% of respondents expected to focus budget on "Customer data integration"

44% expected to focus budget on "Social media monitoring software" [5]

Future focus is on improving online customer experience.

77% of respondents stated "Improving online Customer Experience" as major objective for 2012 (Fig. 5).

Fig.5. Customer Relationship Management

74% stated "improving cross-channel customer experience" as a major objective

As we grapple with the consumption challenges presented by this deluge of data, new publishing platforms are also empowering us to gather, refine, analyze and share data ourselves, turning it into information.

6. Future of Big Data

Clearly Big Data is in its beginnings, and is much more to be discovered. Now is for the most companies just a cool keyword, because it has a great potential and not many truly know what all is about.

A clear sign that there is more to big data then is currently shown on the market, is that the big software companies do not have, or do not present their Big Data solutions, and those that have like Google, does not use it in ca commercial way.

The companies need to decide what kind of strategy use to implement Big Data. They could use a more *revolutionary* approach and move all the data to the new Big Data environment, and all the reporting, modeling and interrogation will be executed using the new business intelligence based on Big Data. [1]

This approach is already used by many analytics driven organizations that puts all the data on the Hadoop environment and build business intelligence solutions on top of it.

Another approach is the *evolutionary* approach; Big Data becomes an input to the current BI platform. The data is accumulated and analyzed using structured and unstructured tools, and the results are sent to the data warehouse. Standard modeling and reporting tools now have access to social media sentiments, usage records, and other processed Big Data items. [1] One of the issues of the *evolutionary* approach is that even if it gets most of the capabilities of the Big Data environment, but also gets most of the problems of the classic Business intelligence solution, and in some cases can create a bottleneck between information that came from the Big Data and the power to analyze of the traditional BI or data warehouse solution

Another approach is the *hybrid* one, where some types of data are analyzed by the Big Data and other by the traditional BI Solutions.

One of the solutions that are now available is the Hana solution from SAP.

The work done in the real-time analytics, in-memory database, with memory becoming cheaper, multi-core architecture which allow parallel processing, compression techniques useful to keep more data in less memory, column and row storage being able to access data at amazing speed and real-time replication capability of data is the role of SAP HANA. These capabilities are illustrated in Fig. 6. [6]

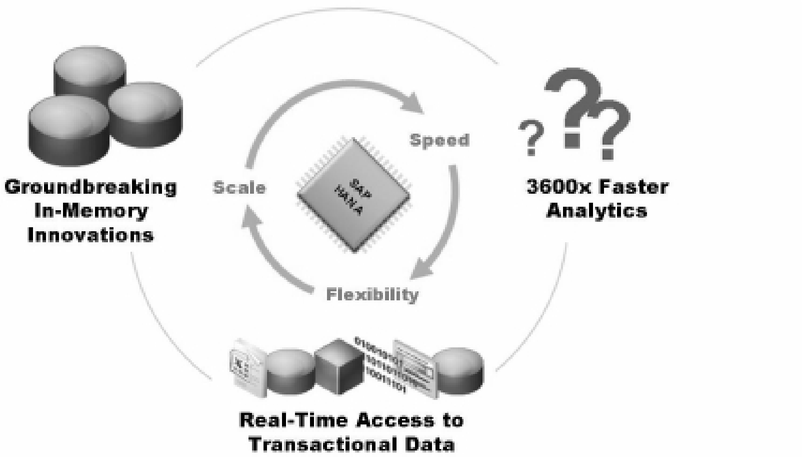

Groundbreaking In-Memory Innovations
Scale
Speed
SAP HANA
Flexibility
3600x Faster Analytics
Real-Time Access to Transactional Data

Fig. 6. The key capabilities of SAP HANA [6]

This technology itself brings business benefits by being leveraged across domains like Big Data, Business Intelligence and Analytics. These business benefits are:

- Speed and Accelerated performance: good query performance for improved decision making, boost of performance for data load processes for a low data latency, accelerated memory planning capabilities [7];
- New Business Insights: Self-Service BI and more flexible modeling capabilities;
- Faster Business Processes;
- Simpler Business Interaction empowering the decision maker to take action in a very short time [7].

SAP HANA as a database would bring performance with the use of in-memory database functionality. The role of SAP HANA in the future of big data is to move all the performance processing planning functions into the in-memory database. This enables planners to:

- Use more data for planning, in this case planning runs can be done daily or for multiple years [12];
- Plan at a detailed level of granularity. For example: Demand for products can be requested at a material and variant level than a product group level. Integrate planning activities across functions [12].

As challenges of large volumes of data appeared and more data in systems are requested, in-memory feature address some business drivers like:

- Manage big data and complexity in an efficient way [6];
- Remove constraints when analyzing big data - trends, data

mining or predictive analytics;

- Allow simulation capabilities for different solutions to do the best choose [6].

Accordingly, big data can be efficient optimized with the help of this revolutionary SAP Hana. One of the main reasons that SAP Hana appeared is big data requirements, which means large transaction volume.

7. Conclusions

In the present Big Data is at the beginning, is very rarely implemented in companies, most of them are getting there even if they don’t realized it yet.

There are many definitions and visions about big data and there is no completely accepted one that everybody can agree on.

As explain earlier, the approach of Gartners Company is the mod popular, but not everyone agree with it because is more business orientated and less visionary.

Many imagine the future of big data like the central nervous system of the planet, with, for which the people are its sensors.

One thing all agree, the Big data concept is one that will revolutionize all businesses, because of the information it holds and the capability to interpret and analyze even the most volatile and non-related data.

I think that Big Data is the solution for many of the world problems, and is now been born to the most productive time in our history. Now is the time when the technology is developing so fast that even a super computer like the ones that are used by Google or Amazon, can be loan or bough at a lesser price than before.

Another advantage of the today technology level is the information that is free on the internet, or the multitude of sensors or companies that have those sensors that can capture valuable information about almost everything that is happening in the world.

The most important issue today is represented by the small number of Big Data specialists. There are very few courses and trainings available in the market, most of them are hold between close doors and most of them are very expensive.

Another issue in the process of training a specialist is the level of knowledge and know-how that is required. So there will be no IT interns learning about Big Data very soon.

In any case I think that Big Data is the next big thing, not only in IT market, but in the life an activity of every company.

References

[1] Dr. Arvind Sathi, *Big Data Analytics*, Ed. Distributive Technologies for Changing the Game, USA 2012.

[2] Judith Hurwitz, Alana Nugent, Dr. Fern Halper, Marcia Kaufman, *Big Data for Dummies, John Wiley & Sons Inc,* USA 2013

[3] Michael Driscoll *Big Data Now*, Ed. O’Reilly, USA 2012.

[4] Pete Warden, *Big Data Glossary*, Ed. O’Reilly, USA 2012.

[5] Rick Smolan, Jennifer Erwit, *The Human face of Big Data*, Ed. Against all odds production, Sausalito, CA 2012.

[6] SAP AG or an SAP affiliate company, SAP HANA Introduction, Participant Handbook, 2013

[7] SAP AG or an SAP affiliate company, SAP HANA Implementation and Modeling, Participant Handbook, 2013.

[8] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E Gruber, *Bigtable: A distributed Storage System for Structured Data,* Google Inc, http://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf

[9] http://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/

[10] https://cassandra.apache.org/#tab-presentation

[11] http://en.wikipedia.org/wiki/MapReduce

[12] DC READINESS SAP HANA. Building a Trusted SAP HANA Data

Center. Internet: http://www.cisco.com/assets/events/i/sapte-hana_whitepaper.pdf, September 2013.

Mircea Răducu TRIFU graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2011 and the Informatics Security Master in 2013. He also finished the Faculty of Management in 2009, and the Master in Business Administration in 2011, both at the Bucharest University of Economic Studies. At present he is a System Support at Data warehouse team in the department of the Application Development of the ING Bucharest.

Mihaela Laura IVAN graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2011. She also finished the Master's degree in Economic Informatics in 2013, at the Bucharest University of Economic Studies. At the present she is a SAP Development Consultant at SAP Near Shore Center Romania.

Forecasting Final Energy Consumption using the Centered Moving Average Method and Time Series Analysis

Janina POPEANGĂ, Ion LUNGU
University of Economic Studies, Bucharest, Romania
janina.popeanga@yahoo.com, ion.lungu@ie.ase.ro

The forecasting of energy consumption has become one of the major fields of research in recent years. Accurate energy demand forecasting is essential in energy system operations and planning.

In this paper, we will describe a method to determine the information that is useful for a good forecasting. Further, we adopt the time series modeling approach to model final energy consumption in Romania using previous data of 2010 to 2013. This method is implemented using stored procedures, developed in Oracle PL/SQL programming language.

Finally, the developed model is compared for goodness of fit to the historical data and forecasting accuracy, and results are encouraging, showing that the forecast model is in control and is working correctly.

Keywords: *Forecasting, Energy, Centered Moving Average Method, Time Series, Accuracy*

1 Introduction

The analysis of temporal data and the prediction of future values of time series are among the most important problems that data analysts face in many fields, ranging from finance and economics, to production operations management or telecommunications. [1]

According to NIST/ITL (1997), time series is generally an ordered sequence of values of a variable at equally spaced time intervals.

Energy consumption recorded over a period of time at fixed interval is a classic time series modeling problem, which is generally used for forecasting.

The forecast for the energy consumption and power also is the scientific activity with the main purpose: the forecast for the energy consumption and power based on calculations analysis and based on the interpretation of different dates, so we will obtain a more precise concordance between the estimated consumptions and the one effectively realized. [2]

A forecast is a prediction of some future event(s).

Accurate final energy consumption forecasting is essential in energy system operations such as during startup and shut-down schedules of generating units as well as for fix planning and spot market energy pricing.

Energy consumption forecasting can be divided into three categories [3]:

- *short term forecasting* - predicts the load demand from one day to several weeks. It helps to estimate and allocate resources in a power grid to supply the demand continuously, to prevent overloading and so lead to more economic and secure energy system.
- *medium term forecasting* – provides information for power system planning and operations, predicting the consumption from a month to several years.
- *long term load forecasting* - predicts the final energy consumption from a year up to twenty years and it is mainly for system planning, allowing decision makers of a power supplying company to decide when to build new power plants, transmission and distribution networks.

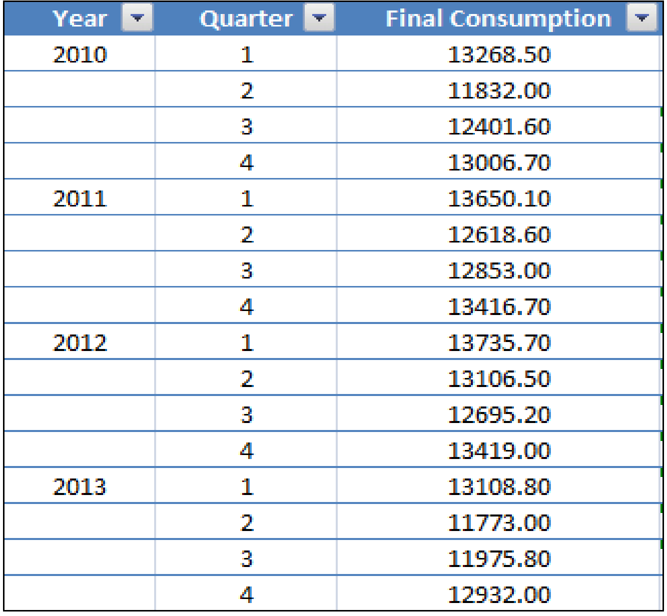
2. Methodological framework

The methodology of elaboration of a forecast study for the energy consumption has few main steps [4]:

- collecting, selection and analyze the initial dates;
- establishing the mathematical model for the consumption;
- the analyze for the variance which has been obtained for the forecast and establishing the final decision.

This paper investigates the effectiveness of a model developed for time series forecasting. Here, we utilize the time series modeling approach to model final energy consumption in Romania using previous data of 2010 to 2013 shown in Table 1.

Table 1. Quarterly Data for Final energy consumption



Year ▾	Quarter ▾	Final Consumption ▾
2010	1	13268.50
	2	11832.00
	3	12401.60
	4	13006.70
2011	1	13650.10
	2	12618.60
	3	12853.00
	4	13416.70
2012	1	13735.70
	2	13106.50
	3	12695.20
	4	13419.00
2013	1	13108.80
	2	11773.00
	3	11975.80
	4	12932.00

Any time series can contain some or all of the following components:
1. *Trend (T)* - Is the long term pattern of a time series;
2. *Cyclical (C)* – up and down movement repeating over long time frame;
3. *Seasonal (S)* - Seasonality occurs when the time series exhibits regular fluctuations during the same month (or months) every year, or during the same quarter every year;
4. *Irregular (I)* - In prediction, the objective is to “model" all the components to the point that the only component that remains unexplained is the random component.

LINEAR TREND EQUATION:

$$Y' = a + b*t$$

where:
Y' is the projected value of the Y variable for a selected value of t.
a is the Y-intercept. It is the estimated value of Y when t=0, so a is the estimated value of Y where the line crosses the Y-axis when t is zero.
b is the slope of the line, or the average change in Y' for each change of one unit in t.
t is any value of time that is selected.

In contrast to the least squares method, which expresses the trend in terms of a mathematical equation ($Y' = a + b$*t), the moving-average method “smooths” the fluctuations in the time series, to see its trend.
When calculating a moving average, placing the average in the middle time period makes sense. But, giving the fact that our data are quarterly, and since there are four quarters in a year, technically, the MA(4) would fall at *t* = 2,5; 3,5; ....
The first value that can be calculated for this series by a 4-period MA process would use observations X_1, X_2, X_3, and X_4. So, first 4-period average has a center between quarter 2 and quarter 3.
Thus we have:

$$X_{2,5} = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

The second MA values would use observations X_2, X_3, X_4 and X_5. So, second 4-period average has a center between quarter 3 and quarter 4:

$$X_{3,5} = \frac{X_2 + X_3 + X_4 + X_5}{4}$$

For the time series, the general formula is:

$$X_{t,5} = \frac{X_{t-1} + X_t + X_{t+1} + X_{t+2}}{4} \quad (1)$$

To avoid this problem, when we average an even number of values, we need to smooth the smoothed values. This method it is called *Double Moving Average for a Linear Trend Process*.

To get a 4-period double moving average that is centered at quarter 3 we take the average of $X_{2,5}$ and $X_{3,5}$:

$$X^{CMA}{}_3 = \frac{X_{2,5} + X_{3,5}}{2}$$

The general formula is:

$$X^{CMA}{}_t = \frac{X_{(t-1),5} + X_{(t+1),5}}{2} \quad (2)$$

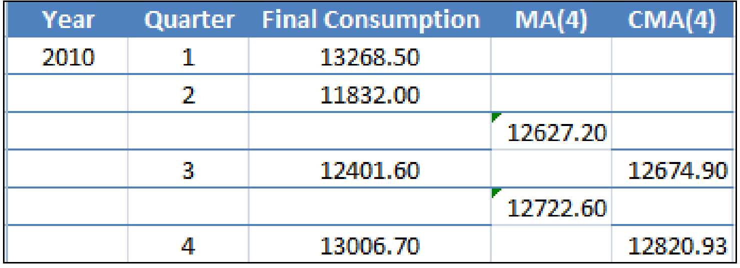
From (1) and (2), results that:

$$X^{CMA}{}_t = \frac{1}{2} * \left(\frac{X_{t-2} + X_{t-1} + X_t + X_{t+1}}{4} + \frac{X_{t-1} + X_t + X_{t+1} + X_{t+2}}{4} \right)$$

=>

$$X^{CMA}{}_t = \frac{X_{t-2} + 2 * X_{t-1} + 2 * X_t + 2 * X_{t+1} + X_{t+2}}{8}$$

Table 2. Calculating *MA*(4) and *CMA*(4)



Year	Quarter	Final Consumption	MA(4)	CMA(4)
2010	1	13268.50		
	2	11832.00		
			12627.20	
	3	12401.60		12674.90
			12722.60	
	4	13006.70		12820.93

The specific seasonal for each quarter is calculated by dividing final consumption in column 3 by the centered moving average in column 5. The specific seasonal reports the ratio of the original time series value to the moving average. Algebraically, we compute final consumption/*CMA*(4) = *SI* and this result is the seasonal component.

Next we calculate a typical seasonal index for the corresponding quarters:

$$S_{t(Q1)} = \text{AVG}\ (SI_{(2011,Q1)}, SI_{(2012,Q1)}, SI_{(2013,Q1)})$$

$$S_{t(Q2)} = \text{AVG}\ (SI_{(2011,Q2)}, SI_{(2012,Q2)}, SI_{(2013,Q2)})$$

$$S_{t(Q3)} = \text{AVG}\ (SI_{(2010,Q3)}, SI_{(2011,Q3)}, SI_{(2012,Q3)})$$

$$S_{t(Q4)} = \text{AVG}\ (SI_{(2010,Q4)}, SI_{(2011,Q4)}, SI_{(2012,Q4)})$$

The set of typical indexes is very useful in adjusting the time series, for example, for seasonal fluctuations. Deseasonalizing the final consumption series is to remove the seasonal fluctuations so that the trend can be studied.

To remove the effect of seasonal variation, the final energy consumption for each quarter is divided by the seasonal index for that quarter:

$$Deseasonalize = final\ consumption\ /\ S_t$$

Next we determine the regression equation of the trend data and use it to forecast future energy consumptions.

$$Regression\ (Deseasonalize,\ t)$$

The trend equation is:

$$Trend_t = Intercept + Slope * t$$

$$Forecast_t = S_t * T_t$$

3. Implementing the method

In order to implement the method described, we create the following tables:

- **tb_FCons**

```
([t] [int] IDENTITY(1,1) NOT
NULL,
[Year] [int] NOT NULL,
[Quarter] [int] NOT NULL,
[Final_cons] [decimal](10, 2) NOT
NULL,
[CMA] [decimal](10, 2),
[SIt] [decimal](10, 2),
[St] [decimal](10, 2),
[Deseasonalize] [decimal](10, 2),
[Tt] [decimal](10, 2),
[Forecast] [decimal](10, 2))
```

- **tb_Regr**

```
([n] [int] NOT NULL,
[slope] [decimal](10, 2),
[intercept] [decimal](10, 2))
```

tb_FCons contains quarterly information on final energy consumption in Romania using previous data of 2010 to 2013.

tb_Regr table stores data about the indicators calculated by using the Linear Regression technique (intercept and slope), for n number of observations.

The main procedures that are used in the application are:

- *QUARTERLY_ANALYSIS* - for calculating the regression indicators and forecast actual values of the time series, in order to measure the accuracy of the model.
- *FORECASTING* – for forecasting final energy consumption for the following 'nQ' quarters (nQ – number of quarters, given as parameter).

All of them are shown below.

```
CREATE PROCEDURE
QUARTERLY_ANALYSIS
AS

DECLARE
      @nr int,
      @Xtim2 decimal(10,2),
      @Xtim1 decimal(10,2),
      @Xti decimal(10,2),
      @Xtip1 decimal(10,2),
      @Xtip2 decimal(10,2),
      @cma decimal(10,2),
      @SIt decimal(10,2),
      @StQ decimal(10,2),
      @Slope decimal(10,2),
      @Intercept decimal(10,2),
      @ti int =3,
      @i int =1

SELECT @nr=count(*) FROM tb_FCons;

WHILE @ti<@nr-1
    BEGIN
        SELECT @Xtim2=Final_cons
FROM tb_FCons WHERE t=@ti-2;
        SELECT @Xtim1=Final_cons
FROM tb_FCons WHERE t=@ti-1;
        SELECT @Xti=Final_cons FROM
tb_FCons WHERE t=@ti;
        SELECT @Xtip1=Final_cons
FROM tb_FCons WHERE t=@ti+1;
        SELECT @Xtip2=Final_cons
FROM tb_FCons WHERE t=@ti+2;

        SET @cma = (@Xtim2 +
2*@Xtim1 + 2*@Xti + 2*@Xtip1 +
@Xtip2)/8;
        SET @SIt = @Xti/@cma;

        UPDATE tb_FCons
        SET CMA=@cma, SIt=@SIt
        WHERE t=@ti;

        SET @ti=@ti+1;

      END

WHILE @i<=4
      BEGIN
         SELECT @StQ = avg(SIt)
         FROM tb_FCons
         WHERE Quarter=@i
         GROUP BY Quarter;

         update tb_FCons
         SET St=@StQ
         WHERE Quarter=@i;

         SET @i=@i+1;
      END

UPDATE tb_FCons
SET Deseasonalize=Final_cons/St;

SELECT
      @Slope = ((@nr *
sum(t*Deseasonalize)) -
(sum(t)*sum(Deseasonalize)))/ ((@nr
* sum(Power(t,2)))-Power(Sum(t),2)),
      @Intercept = avg(Final_cons) -
((@nr * sum(t*Deseasonalize)) -
(sum(t)*sum(Deseasonalize)))/((@n
r * sum(Power(t,2)))-
Power(Sum(t),2)) * avg(t)
FROM tb_FCons

UPDATE tb_FCons
SET Tt = @Intercept + @Slope*t,
    Forecast = St*Tt;

INSERT INTO tb_Regr VALUES
(@nr,@Slope, @Intercept);
```

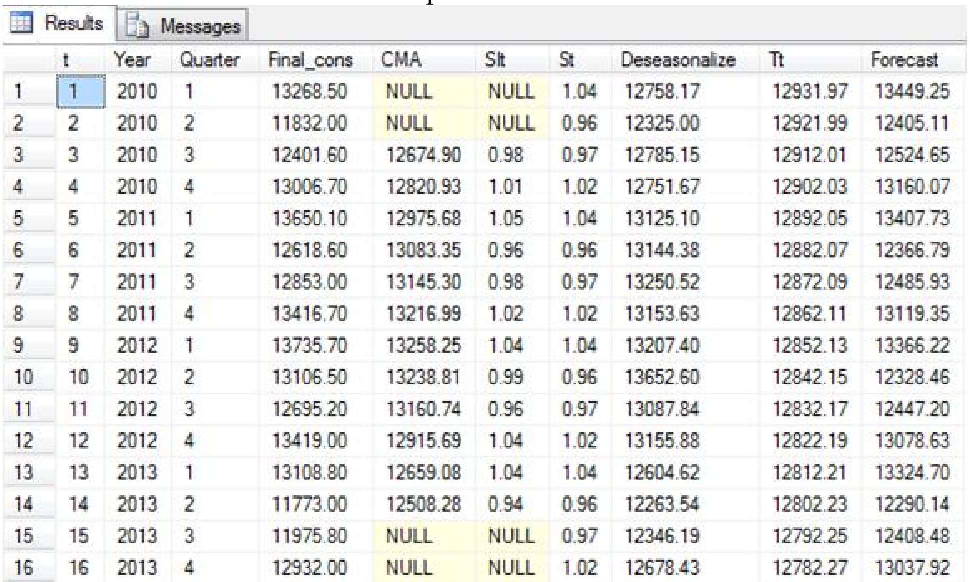
The second algorithm was implemented by creating the temporary table *#Temp_FCons* that will capture the forecasting final energy consumption values for the period desired.

```
CREATE PROCEDURE FORECASTING
          @nQ int
AS

DECLARE
    @t int,
      @nr int,
      @y int,
      @q int,
      @StQ decimal(10,2),
      @Slope decimal(10,2),
      @Intercept decimal(10,2);

IF
object_id('tempdb..#Temp_FCons')
is not null DROP TABLE
#Temp_FCons

CREATE TABLE #Temp_FCons(
      [t] [int] NOT NULL,
      [Year] [int] NOT NULL,
      [Quarter] [int] NOT NULL,
      [St] [decimal](10, 2),
      [Tt] [decimal](10, 2),
      [Forecast] [decimal](10, 2))

SELECT @t=count(*)+1 FROM tb_FCons;
SELECT @y=max(YEAR) FROM tb_FCons;
SET @nr=@t+@nQ;

WHILE @t<=@nr-1

BEGIN

IF @t% 4=1 SELECT @y=@y+1, @q=1
ELSE IF  @t% 4=0 SELECT @y=@y, @q=4
      ELSE SELECT @y=@y, @q=@t % 4

SELECT @StQ = avg(SIt)
FROM tb_FCons
WHERE Quarter=@q
GROUP BY Quarter;

SELECT @Slope= slope,
       @intercept = intercept
FROM tb_Regr;

INSERT INTO #Temp_FCons (t, Year,
Quarter,St,Tt) VALUES
(@t,@y,@q,@StQ,@Intercept +
@Slope*@t);

SET @t=@t+1;
  END

  UPDATE #Temp_FCons
  SET Forecast = St*Tt;

  SELECT * FROM #Temp_FCons
```

4. Results and analysis

The execution of *QUARTERLY_ANALYSIS* procedure will produce the result set shown in *Table 2*.

Table 2. Forecasting final energy consumption using the QUARTERLY_ANALYSIS stored procedure



	t	Year	Quarter	Final_cons	CMA	Sit	St	Deseasonalize	Tt	Forecast
1	1	2010	1	13268.50	NULL	NULL	1.04	12758.17	12931.97	13449.25
2	2	2010	2	11832.00	NULL	NULL	0.96	12325.00	12921.99	12405.11
3	3	2010	3	12401.60	12674.90	0.98	0.97	12785.15	12912.01	12524.65
4	4	2010	4	13006.70	12820.93	1.01	1.02	12751.67	12902.03	13160.07
5	5	2011	1	13650.10	12975.68	1.05	1.04	13125.10	12892.05	13407.73
6	6	2011	2	12618.60	13083.35	0.96	0.96	13144.38	12882.07	12366.79
7	7	2011	3	12853.00	13145.30	0.98	0.97	13250.52	12872.09	12485.93
8	8	2011	4	13416.70	13216.99	1.02	1.02	13153.63	12862.11	13119.35
9	9	2012	1	13735.70	13258.25	1.04	1.04	13207.40	12852.13	13366.22
10	10	2012	2	13106.50	13238.81	0.99	0.96	13652.60	12842.15	12328.46
11	11	2012	3	12695.20	13160.74	0.96	0.97	13087.84	12832.17	12447.20
12	12	2012	4	13419.00	12915.69	1.04	1.02	13155.88	12822.19	13078.63
13	13	2013	1	13108.80	12659.08	1.04	1.04	12604.62	12812.21	13324.70
14	14	2013	2	11773.00	12508.28	0.94	0.96	12263.54	12802.23	12290.14
15	15	2013	3	11975.80	NULL	NULL	0.97	12346.19	12792.25	12408.48
16	16	2013	4	12932.00	NULL	NULL	1.02	12678.43	12782.27	13037.92

Fig.1 compares actual final energy consumption histories to forecasts for the entire time period analyzed.

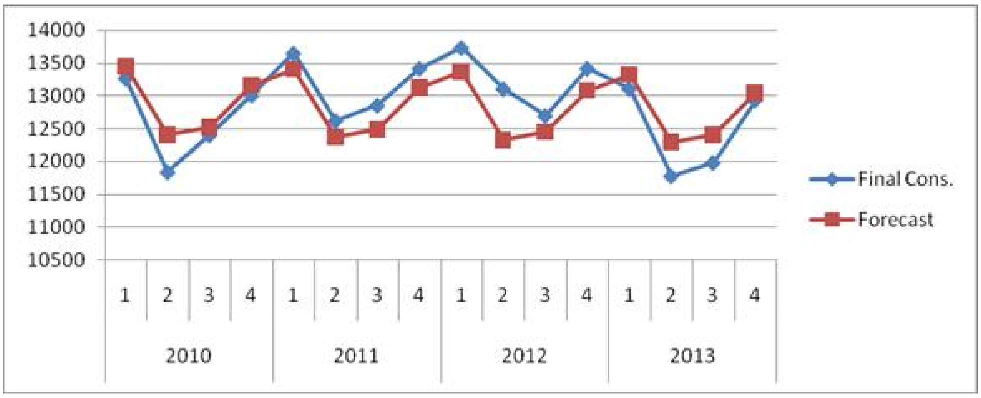


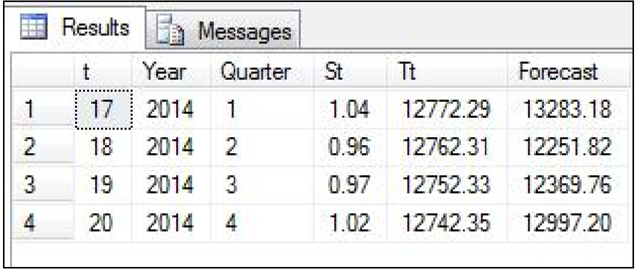
Fig. 1 – Actual and Forecast Final energy consumption (mill. kWh)

Further, we can call this store procedure using EXEC, and specifically specifying the parameter of the procedure:

```
EXEC FORECASTING @nQ=4
```

will display the results for the future 4 quarters (*Table 3*).

Table 3. Forecasting results



	t	Year	Quarter	St	Tt	Forecast
1	17	2014	1	1.04	12772.29	13283.18
2	18	2014	2	0.96	12762.31	12251.82
3	19	2014	3	0.97	12752.33	12369.76
4	20	2014	4	1.02	12742.35	12997.20

According to our analysis, the final energy consumption is expected to increase around 3% in the first quarter and then decrease by 7,77% in quarter Q2. Total energy consumption will increase from 12251,82 million kWh in Q2 to

12369,76 million kWh in Q3 and 12997,20 million kWh in Q4. (*Fig. 2*)

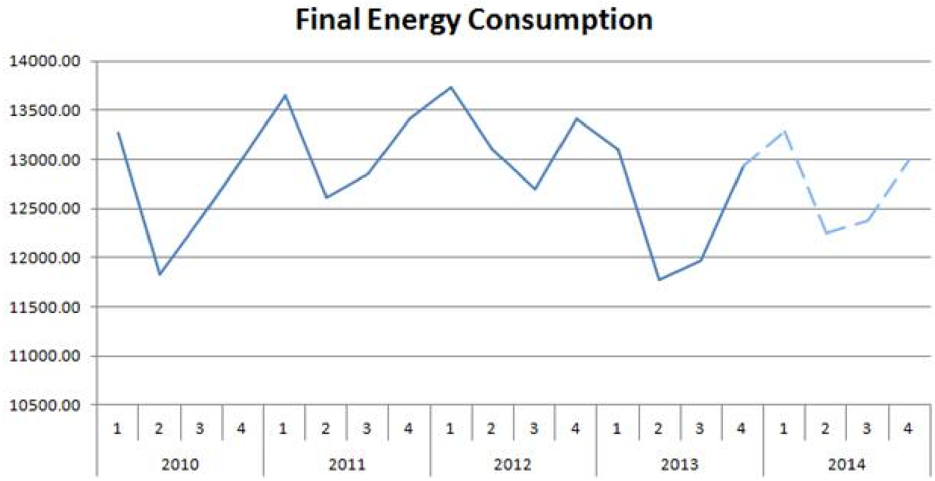


Fig. 2 – Final energy consumption (mill. kWh)

5. Evaluation of Forecasting Model

Forecast error = Difference between actual and forecasted value (also known as residual).

$$Forecast\ error = Actual - Forecast$$

The Mean Absolute Deviation (MAD) is calculated by adding the absolute value of forecast errors in each period, and taking the average of this total.

$$MAD = \frac{\sum_{t=1}^{n} |A_t - F_t|}{n}$$

The Mean Squared Error (MSE) measures the average of the squares of the errors, that is, the difference between the actual energy consumption and what is estimated.

$$MSE = \frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}$$

The Tracking Signal indicates if the forecast is consistently biased high or low and is calculated as the ratio of cumulative error and MAD.

$$Tracking\ Signal = \frac{\sum_{t=1}^{n} A_t - F_t}{MAD}$$

Tracking signal values are compared to predetermined limits (+4,-4) based on judgment and experience.

If TS > 4 or < -4 => investigate!

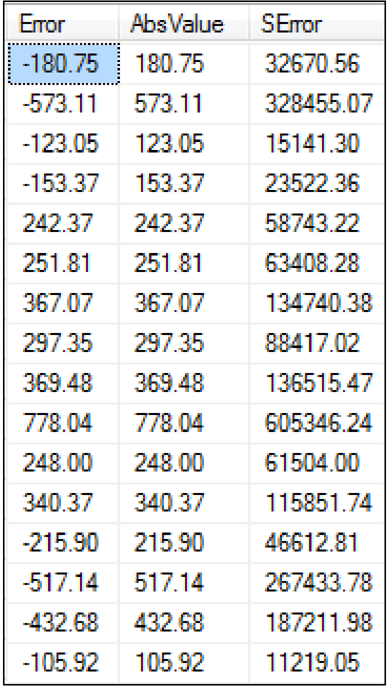
If TS ≥ 0 => most of the time actual values are above the forecasted values.

If TS < 0 => most of the time actual values are below the forecasted values.

We investigate our forecasting model by querying the tb_FCons table:

```
SELECT (Final_cons - Forecast) as
Error, abs(Final_cons - Forecast) as
AbsValue, power((Final_cons -
Forecast),2) as SError
FROM tb_FCons
```

Table 4. Forecasting accuracy



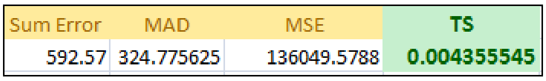
Error	AbsValue	SError
-180.75	180.75	32670.56
-573.11	573.11	328455.07
-123.05	123.05	15141.30
-153.37	153.37	23522.36
242.37	242.37	58743.22
251.81	251.81	63408.28
367.07	367.07	134740.38
297.35	297.35	88417.02
369.48	369.48	136515.47
778.04	778.04	605346.24
248.00	248.00	61504.00
340.37	340.37	115851.74
-215.90	215.90	46612.81
-517.14	517.14	267433.78
-432.68	432.68	187211.98
-105.92	105.92	11219.05

Further, we calculate the following indicators using the accuracy results obtained at the previous step:

$$MAD = \text{AVG}\ (AbsValue)$$

$$MSE = \text{AVG}(SError)$$

$$TS = \text{SUM}(Error)\ /\ MAD$$



Sum Error	MAD	MSE	TS
592.57	324.775625	136049.5788	0.004355545

Model tends to over-forecast, with an average absolute error of 324. 77 units.

TS control limits of ±2 to ±4 are used most frequently. Values outside this rage indicate that the model should be investigated and reevaluated.

Therefore, our example shows that this forecast is in control and this model is working correctly.

5. Conclusions

Forecasting the energy consumption is the scientific activity with the main objective: to obtain a more precise concordance between the estimated consumptions and the one effectively realized, based on calculations analysis and based on the interpretation of different dates.

In the first part of our investigation, we have collected, selected and analyzed the initial dates. The role of the Double Moving Average for a Linear Trend Process is explained and detailed.

Second, we have established the mathematical method for calculating the expected energy consumption values;

This method is implemented through stored procedures that calculate the regression indicators, forecast actual values of the time series, in order to measure the accuracy of the model and forecast final energy consumption for the following quarters.

Evaluation of forecasting model shows that the model developed works correctly.

References

[1] Martinez Alvarez, F. Troncoso, A., Riquelme, J.C.; Aguilar Ruiz, J.S., *Energy Time Series Forecasting Based on Pattern Sequence Similarity*, Knowledge and Data Engineering, IEEE Transactions on (Volume:23 , Issue: 8), 2011.

[2] Hamilton, J. “Time series analysis”, Princeton University Press, 1994.

[3] Intan Azmira binti Wan Abdul Razak, Shah bin Majid, Mohd Shahrieel bin Mohd. Aras and Arfah binti Ahmad, “Advances in Data Mining Knowledge Discovery and Applications”, Chapter 11 - Electricity Load Forecasting Using Data Mining Technique, ISBN 978-953-51-0748-4, Published: September 12, 2012

[4] P. Mihai, M. Popescu, “Mathematical Models used in Quality Management of the Electrical Energy”

Janina POPEANGĂ graduated in 2010 the Faculty of Cybernetics, Statistics and Economic Informatics, Economic Informatics specialization. The title of her Bachelor’s thesis is “*Distributed Databases*”. In 2012, she graduated the Databases for Business Support master program with the thesis “*Monitoring and management of electric power consumption using sensorial data*”. Janina’s interests are broadly in the fields of databases and distributed systems. Since 2012 she is a Ph.D. Student in the Doctoral School of Bucharest Academy of Economic Studies. Her research focuses on real-time database systems, business intelligence analytics, sensor data management, smart grid and renewable energy.

Ion LUNGU graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 1974. He got the title of doctor in economy in the specialty economic informatics in 1983. He has been directing graduates who study towards getting a doctor’s degree since 1999. At present he is a professor in the department of the faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies of Bucharest. He had documentary activity and specialization with the Eindhoven Technical University of Holland, the Economic University of Athens and Economic University of Milan. His domains of work are: informatics systems and databases. Among his books are: "Databases, organization, design and implementation", (1995), "Information Systems for Management" (1994), "SGBD Oracle Applications" (1998); "Let’s learn Oracle in 28 lessons” (2003), "Database systems" (2003), "Information Systems – Analysis, Design and Implementation" (2003).