Information Systems & Grid Technologies

Eighth International Conference ISGT'2014

Sofia, Bulgaria, 30 – 31. May, 2014

ISGT'2014 Conference Committees

Chair

Prof Vladimir DIMITROV

Program Committee

- Míchéal Mac an AIRCHINNIGH, Trinity College, University of Dublin
- Pavel AZALOV, Pennsylvania State University
- Marat BIKTIMIROV, Joint Supercomputer Center, Russian Academy of Sciences
- Marko BONAČ, Academic and Research Network of Slovenia
- Marco DE MARCO, Catholic University of Milan
- Milena DOBREVA, University of Strathclyde, Glasgow
- Viacheslav ILIN, Moscow State University
- Vladimir GETOV, University of Westminster
- Jan GRUNTORÁD, Czech Research and Academic Network
- Pavol HORVATH, Slovak University of Technology
- Seifedine KADRY, American University of the Middle East, Kuwait
- Arto KARILA, Helsinki University of Technology
- Dieter KRANZMUELLER, University of Vienna
- Shy KUTTEN, Israel Institute of Technology, Haifa
- Vasilis MAGLARIS, National Technical University of Athens
- Violeta MANEVSKA, University "St. Kliment Ohridski" – Bitola
- Dov TE'ENI, Tel-Aviv University
- Stanislaw WRYCZA, University of Gdansk
- Fani ZLATAROVA, Elizabethtown College

Organizing Committee

- Vladimir DIMITROV
- Maria NISHEVA
- Kalinka KALOYANOVA
- Vasil GEORGIEV

Vladimir Dimitrov, Vasil Georgiev (Editors)

Information Systems & Grid Technologies

Eighth International Conference ISGT'2014

Sofia, Bulgaria, May, 30 – 31., 2014.

Proceedings

organized by

Faculty on Mathematics and Informatics.
University of Sofia St. Kliment Ohridski

Bulgarian Chapter of the
Association for Information Systems (BulAIS)

St. Kliment Ohridski University Press

Preface

This conference was being held for the eighth time in the end of May, 2014 in the halls of the Faculty of Mathematics and Informatics of the University of Sofia "St. Kliment Ohridski", Bulgaria. It is supported by the Science Fund of the University of Sofia "St. Kliment Ohridski" and by the Bulgarian Chapter of the Association for Information Systems (BulAIS). Traditionally this conference is organized in cooperation with the Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences.

Total number of papers submitted for participation in ISGT'2013 was 31. They undergo the due selection by at least two members of the Program Committee. This book comprises 18 papers of 19 Bulgarian and 14 foreign authors. The conference papers are available also on the ISGT web page http://isgt.fmi.uni-sofia.bg/ (under «Former ISGTs» tab).

The editors

ISSN 1314-4855
St. Kliment Ohridski University Press

TABLE OF CONTENTS

Modeling the Neural Network - Perceptron to Estimate the Safety of the Enterprises in the Republic of Macedonia
Viktorija Stojkovski, Kostandina Veljanovska 7

Reengineering of Software Processes in Municipality for Construction Permission Requests
Snezana Savoska, Branko Dimeski 14

EPC – better option for business process generating
Ivaylo Kamenarov, Katalina Grigorova 25

Cloud Technologies Application at JINR
Nikita Balashov, Alexandr Baranov, Nikolay Kutovskiy, Roman Semenov 32

Hierarchy and expressions for automated workflows for NGS data processing
Milko Krachunov 38

Re Pub(lic) of Philo(sophy)
Mícheál Mac an Airchinnigh, Vassil Nikolov 49

Large Scale Analytics with Hadoop
Vladimir Dimitrov 58

Distributed Coordination with Apache ZooKeeper
Daniel Simeonov, Vasil Georgiev 65

An Ontology-Based Approach for Integrating of Clinical and Molecular Information for Assistance in Medical Diagnostics
Dimitar Vassilev, Maria Nisheva, Nikola Ranchev, Velko Ilchev 74

Methods and Technologies for Email Protection
Falak Hasan 81

The use of the m-banking in the Republic of Macedonia
Marina Blazekovic, Viktorija Stojkovski, Monika Angeloska-Dichovska 93

Architecting Cloud Super Layer of Open Source Components
Hristo Hristov, Vasil Georgiev 100

Measuring Influence of Genome Annotation Version to Data Analysis Results
Ognyan Kulev 105

Business Processes in Grid and Cloud
Radoslava Hristova, Vladimir Dimitrov 115

Business Process Model Based on Business Rules
Evgeniy Krastev, Maria Semerdjieva 122

The Design and Realization of the Municipal Informational and Administrative Website
Krasimir Nikolov, Svetlana Vasileva 133

Analysis of Business Process Models
Kristiyan Shahinyan, Evgeniy Krastev 144

Distributed Training and Testing Grid Infrastructure Evolution
Radoslava Hristova, Nikolay Kutovskiy, Vladimir Dimitrov, Vladimir Korenkov 158

AUTHOR INDEX 163

Modeling the Neural Network - Perceptron to Estimate the Safety of the Enterprises in the Republic of Macedonia

Viktorija Stojkovski, Kostandina Veljanovska

Faculty of Administration and Information Systems Management, Univetrsity " St. Kliment Ohridski", Partizanska bb, 7000 Bitola, Republic of Macedonia

Abstract. Neural networks are a metaphor of the human brain used for information processing. It is shown that they are very promising techniques for various applications and classified business applications by their ability to "learn" from the data, their nonparametric nature and also, their ability to generalize. The aim of this paper is to model the perceptron where using properly selected inputs and their weights we obtain a suitable output. Learning rule of the perceptron understands that the model is developed in order to train the perceptron, ie. the learning algorithm is developed where weights are adjusted to minimize the error when the network output does not corresponds to the desired output. The measures used by businesses to protect themselves and also their exposure to the Internet, which is a potential threat to their security are used as input data. The expected outcome of the paper are the analysis of the suitability of the perceptron in estimating the level of security for businesses in terms of the use of information and communication technologies during the phase of model's initial experiments.

Keywords: Perceptron, Neural Network, IT Security , Learning Rule

1 Introduction

In fact, the neural network is a structured model with algorithm for "feeding" the model. The result from "neural computers" is a model which is called artificial neural network or neural network. Neural networks are used for many business application like recognition, planning, prediction and classification. Neural network architectures are: feedforward network and feedback network. Further, feedforward networks are divided into: one layer perceptron, multilayer perceptron, radial basic function network, and feedback networks are divided into: Competitive, Kohonen's SOM, Hopfield and ART.

2 Perceptron

Perceptron's history started in late 1950, when Frank Rosenblatt and few other researchers developed class of neural network named perceptron. Neurons

in that network were similar to neural network which is developed by McCulloch and Pitts, but Rosenblatt's main contribution was introducing the learning rule in order to train perceptron network which later will be used to solve recognized problems. Rosenblatt proved that his learning rule will always converge in accurate weight. Examples are presented to the network to learn the appropriate behavior. Also, the perceptron learns with its weights and bias initialized with random values.

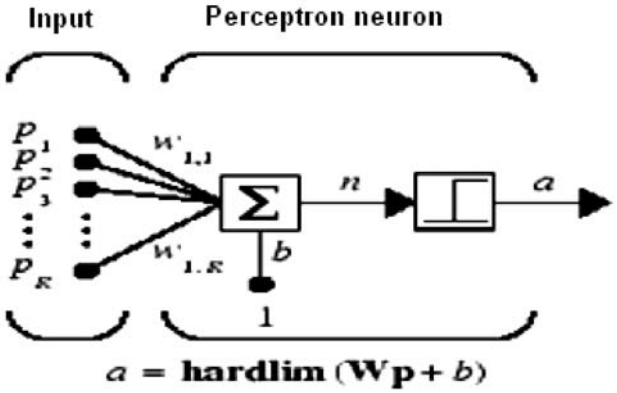

Fig. 1. Perceptron neuron that uses hardlim function

Neural network which contains one input layer with "forward feeding" toward one output layer is known as Perceptron with one layer [3]. Output from the Perceptron with one layer network is shown as:

$$y_j = f(net_j) = \begin{cases} 1 \text{ if } net_j \geq 0 \\ 0 \text{ if } net_j < 0 \end{cases} \quad where\ net_j = \sum_{i=1}^{n} x_i w_{ij}$$

2.1 Perceptron Work Principle

Neuron from perceptron network produces 1, when transfer function input is equal to 0 or more than 0 and produces 0 when it is lower than 0. This function enables Perceptron to classify input vectors dividing the input space into two regions.

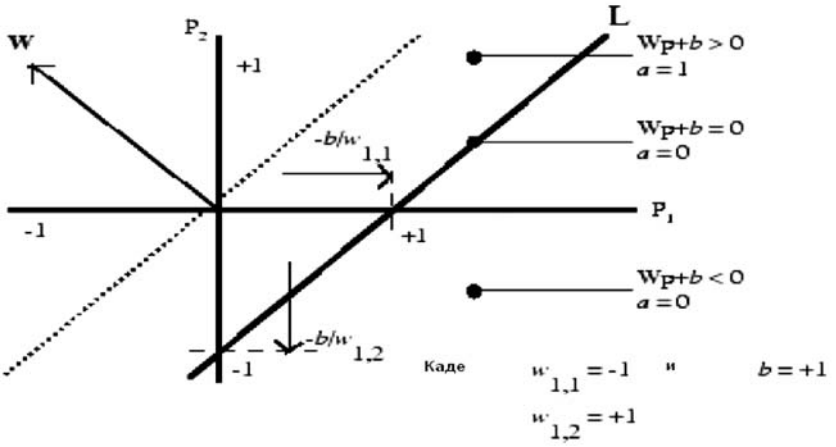

Fig. 2. Input space of two inputs with hardlim

The classification into two regions is shown in Figure 2 with line L at **W**+ b = 0. At the top and left from the line L, input vector will result in network input higher than 0 and for that this hardlimit function will produce output 1 and below and right from the line will produce output 0 respectively. Hard-limit neurons without bias will always classify in line which passes through the original entry. Adding bias enables to solve problems where two sets of input vectors are not located at different sides from the original.

2.2 Perceptron Learning Rule

Perceptron learning rule actually is a procedure for modifying weights and network bias. Also, this procedure is known as algorithm training. The aim of the learning rule is to train the network for reaching the certain goal. In the supervised learning, the learning rule is proved with set of examples of network's adequate behavior as follows:

{p, t}, {p,t},….{p,t},

Where pis input in the network, tis adequate correct output form network [4]. The network inputs, are shown to the network and obtained outputs are compared with desired outputs. When desired outputs are compared with the actual outputs, the learning rule is used to adjust weights and network bias aiming network output to be closer to the desired output. Perceptron learning rule belongs to supervised learning category. Perceptron training rule is an algorithm for learning where weights are adjusted for error minimizing when network output doesn't show the desired output.

- If output is correct, than adjusting weights doesn't exist as follows:

$\{p, t_1\}, \{p_2, t_2\}, \ldots \{p_Q, t_Q\},$

- If output is 1 , but should be 0, then weights are decreased at active entrance as follows:

$$w_{ij}^{k+1} = w_{ij}^{k} - xi$$

- If output is 0, but should be 1, then weights are increased at active entrance as follows:

$$w_{ij}^{k+1} = w_{ij}^{k} + xi,$$ where:

  - w_{ij}^{k} is the new one adjusted weight and w_{ij}^{k+1} is the old weight
  - *xi* is input and is the learning rate
  - low value of means to decrease learning, and higher value of means to increase learning[3].

2.2 Multilayer Perceptron

Multilayer perceptron is the most popular type of neural network used today. Multilayer perceptron structure is divided into layers. The first and the last layer are known and they are input layer and output layer, because they represent inputs and outputs of the whole network. The other layers are hidden layers.

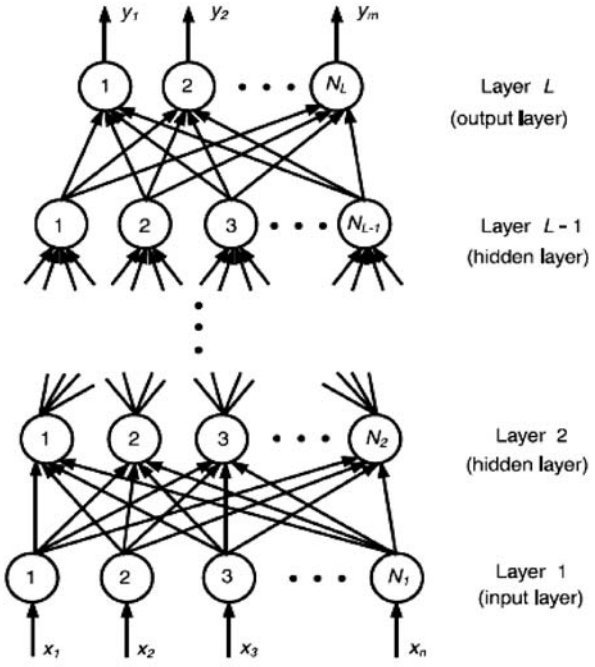

Fig. 3. Multylayer perceptron

Hidden neuron's most used activation function is the sigmoid function trough formula below:

$$\sigma(\gamma) = \frac{1}{(1+e^{-\gamma})}, \text{ where}$$

$$\sigma(\gamma) \to \begin{cases} 1 \text{ as } \gamma \to -\infty \\ 0 \text{ as } \gamma \to -\infty \end{cases}$$

Multilayer perceptron's most used training method is backpropagation or error backpropagation from output to inputs and weights adjustments. Using the backpropagation algorithm, network firstly propagate input from the input layer to the output layer and after the error is determined, it is propagated back to the input layer, but it is embeded in the learning formula. The optimization of the error backpropagated in the network is done using the deterministic algorithm of gradient descent.

3 The Safety of Business Entities

3.1 General View

In order to achieve high security level business entities need to use some technical solutions in combination with corporative security policies. For setting up a security system few steps are needed such as: Setting individual security management, network access control, which includes implementing a firewall, proxy server, system or network administration to set password access control of users and devices on the network, as well as a regular review of network, storage, operating system and also, software regularly installing updates, protection of computers in the company by installing updates, installing antivirus software and firewall regularly involved with the disclaimer to not exclude it.[6]

3.2 Research Subject

Modeling the tool for estimating the level of security for business entities in Republic of Macedonia is the research subject in this paper. Safety is reviewed in terms of the placement of a security system on the one hand and exposure to the Internet on the other hand which could be a threat to their security. The data that will be used as inputs to the network are obtained by adjusting the answers of questionnaire. For the modeling of the single-layer perceptron with two inputs will be used two variables which will be taken into consideration from the questions in the questionnaire "Does the company has an IT sector?" And "Does your employees at their daily operations use the Internet?" Then these variables will be assigned random weights. Modeling the single-layer perceptron with three inputs will be complemented with another input that will be taken into consideration from the question of the questionnaire, "Does your company

has set specific procedures for protection?" Also, the new input is assigned with random weight.

4 Analysis and Discussion of Results

Modeling the single-layer perceptron with two inputs was made with data from a survey which was used as inputs. Input data were: the existence of IT sector in the companies and usage of Internet in daily operations of the company. If companies have IT sector then the input is 1, and if they use the Internet in everyday work input is 0, because of the exposure of the Internet is a potential security threat to businesses. As random weight bias is taken the value 0.1, the value of the first input, the existence of the IT sector is 0.2, and the weight of the second input, the exposure of the Internet is 0.3 and the operator used is AND. The activation function of the model is a function of the threshold. The condition for activating the function is if the sum of the inputs and their weights is greater than 0, then the output is 1, and if it is less than 0, the output is 0. Weights in the model remain the same until the error is obtained by following formula:

$$\text{Error} = \text{Output} - \text{Activation} \quad \textbf{(1)}$$

Perceptron model with two inputs converged in the second epoch after 300 iterations with adjusted weights and weight of bias 0.3, the weight of the first input 0.3 and the weight of second input 0.1.

Table 1. Perceptron model with two inputs

Epoch	Number of iteration	Learning rate	Weight 1	Weight 2	Bias
1	150	0,1	0,2	0,3	0,1
2	150	0,1	0,3	0,1	0,3

Also, perceptron model with three inputs was made, where despite two previous inputs it is used a third input, and that is whether companies have set procedures for protection, which means limited access to confidential information, corporate e-mail and so on. This model converged after 448 iterations in epoch 3, with adjusted weights 0.3 for the first input, for second input 0.2 and 0.1 for the third input.

Table 2. Perceptron model with three inputs

Epoch	Number of iteration	Learning rate	Weight 1	Weight 2	Weight 3	Bias
1	150	0,1	0,1	0,3	-0,2	0,2
2	150	0,1	0,1	0,2	-0,1	0,2
3	150	0,1	0,2	0,1	0,1	0,3

Since both models have been converged, the test was done by inserting other values to the inputs. The network exhibits no errors and gives satisfactory output.

5 Conclusion

Neural networks can be used for many purposes for business applications such as identification, planning, prediction and classification. In particular the modeling of perceptron neural network is made with training of the data that the network learns, their weights and using bias. In modeling perceptron network for assessment of businesses level of security in the Republic of Macedonia the operator AND was used. But, in order to solve more complex problems this operator has limited power. Using another opearator will be the subject of future research. The results and experience gained during the execution of the experiment, shows that to obtain greater precision in evaluation whether the business entity is safe or not, it is necessary to make a model which will use more inputs. Also, a multilayer perceptron network model can be made with multiple layers, using "backpropagation" method of training.

References

1. Jain, K.A., J.Mao, K.M. Mohiuddin: Artificial Neural Network. Available from http:// www.cse.msu.edu
2. Fausett, Lauren V.: Fundamentals of Neural Network, Architectures, Algorithms and applications, Prentice Hall International, pages 60-76, (1994)
3. Jantzen, J: Introduction To Perceptron Networks, http://saba.kntu.ac.ir/eecd/fatehi/Lectures/Intelligent%20Systems/NeuNet/Papers/NeuralNetworksTutorial.pdf
4. Fundamental of Neural Networks, http://www.myreaders.info/08_Neural_Networks.pdf
5. Perceptron Networks, http://www.mathworks.com/help/nnet/ug/perceptron-neural-networks.html
6. Internet Security Essentials for Business 2.0/ www.uschamber.com
7. Stojkovski V, Veljanovska K, "Architectura na bezbednosen system", VNTROPEKOBRM, Zbornik na trudovi, Skopje, 2013, pages.283-292
8. Stojkovski V; E-bezbednost na delovnite subjekti vo Republika Makedonija I bezbednosni aspekti na elektronskata trgovija, Univerzitet Sv. Kliment Ohridski – Bitola, Ekonomski Fakultet –Prilep, (2012)
9. Perceptron Learning Rule, http://hagan.okstate.edu/4_Perceptron.pdf
10. Neuron Model and Network Architectures, http://hagan.okstate.edu/2_Architectures.pdf
11. Veljanovska, K. Advanced Software Techniques, Lecture Notes, UKLO, (2012)
12. Carlo U. Nicola, The Perceptron and its Learning Rule, SGI FH Aargau, http://ebookbrowsee.net/mlpi-pdf-d452908847

Reengineering of Software Processes in Municipality for Construction Permission Requests

Snezana Savoska, Branko Dimeski

Faculty of administration and Information systems Management, University „St.Kliment Ohridski“ – Bitola,
Bitolska bb,
7000 Bitola, R.of Macedonia,
savoskasnezana@gmail.com, branko_dim@yahoo.com

Abstract.. The software processes associated with some administrative procedures sometimes can undergo a process of re-engineering especially when an emerging technology demands a new organization of the whole process. Also, there are some legal changes of the administrative procedures. For that reason, we present the old and transformed process with re-engineering of the software processes within the municipality. These processes support a specific topic which is part of the administrative document issues within the municipal sector of urban and communal planning. Instead of a classic software tools, web based solution have to be made. The whole solution is separated in two main parts: for public citizen service and the appropriate public administration institutions responsible for services. Our research's aim is to gain knowledge of administrative processes that need to be supported by applying software based solutions within municipality in the Republic of Macedonia by the local public administration.

Keywords: Re-engineering, Business processes, Conceptual design, Logical design, Web based information systems, Urban planning.

1 Introduction

When we talk about re-engineering of software processes, we usually talk about finding effective and efficient ways of solving some problems and gaining faster and better solutions (Haddad, 2011). With new emerging technologies, the processes can be simplified and enhanced, their availability increased and also, the risk of errors minimized. A previous version of software processes and structures was connected with a software solution that largely depended on human factor and was centralized within the Department for Urban affairs, Public works, Transport and Environmental protection (More precisely, see the Department of Urban and Public Affair in the text below). The software solution was not possible to be applied anywhere except at the location designed for this purpose. This decentralized software was a major obstacle to the integration of the software into the overall operations of local government. For this reasons, supported by globalization demands, ZELS (The Union of Units of Local Governments) and

the Ministry of Transport and Communications proposed creation of a new web-oriented information system with a different approach in setting up administrative processes. The software solutions have to be modern ones, web oriented and user-friendly (Laplante, 2012). Also, an e-government concept requires strong procedures for supporting concepts that enables web oriented solutions for citizens (McLaughlin, 2007).

In this paper we describe how the old system can be improved and transformed with new processes by re-engeneering. The new system has to be more effective, efficient, time-saving and better organized. With the new software solution, the civil public services can be enhanced and the citizen satisfaction can be improved.

The software processes associated with the old procedures demand better knowledge of complex administrative procedures, given the small number of officials, separated within each local government. For this reason, the absence of officials usually means inability to get things done within the deadline. The new web-based software solution for this administrative problem has to be made already by a Macedonian IT company according to the administrative procedures provided by the existing governmental policy.

2 Legacy Information System used for Issuing Documents for Construction Permission

The case study of a legacy information system for issuing documents for construction permission comes from the Department of Urban and Public Affairs within the municipalities in the Republic of Macedonia. Software solution which they used was made regionally and demanded knowledge of the prescribed procedures. The employees used the software until 2002 and they were able to produce all the needed documents by using software solution and to respond to client's construction permission requests (See below Figure 1).

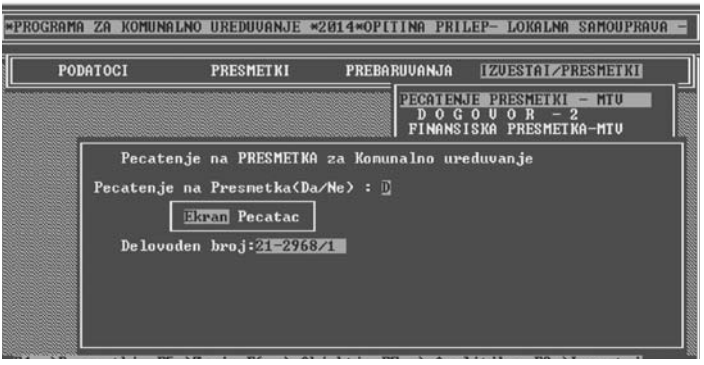

Figure 1 – A screenshot of a legacy system of issuing documents for construction permission requests

The first step was filling administrative application of issuing construction permission request. Also, all needed documentation (as building plans, approval for right to build, proof of payment of administrative fees and the other documents) had to be provided. The software solution enabled a creation of a detail Calculation of construction agreement, Financial and Detail Calculation for construction and issuing an Invoice of a municipal fee. The clerk entered all the necessary data for a classification of urban zones, types of facilities, locations and the other classification issues. Also, they provided building prizes for all zones and types of facilities, detail sizes of the living and business area within the interface and after that they got a precise calculation of the proposed living and business space. Depending on the calculated area, the invoice for municipal fee was created automatically. Software solution also provided reports and searching facilities (See below Figure 2).

But, all these procedures were intended for the clerk in the Sector of Urban and Public Affair (as was shown in Figure 2). There were not any procedures and facilities for clients within the software solution. For that reason, under the mutual project of ZELS (Union of units of local government) and the Ministry of Transport and Communications was decided a provision of an e-service, so-called: "Information system of e-approval and construction". That system had to provide a new approach of the problem solving and also, enhance the existing software solution in some local governments. Also, this software solution had to upgrade the approach of the client as part of e-government solution. The new software solution as e-service had to be web based, user friendly and to provide specific information for the clients and the governmental institutions. They had to provide aggregated data for the Ministry of Transport and Communications officers, for municipal officers and for the government. Also, the solution had to include all the previous operation of local government office support. For that reasons, this part (legacy information system for construction permission requests) had to be reengineered but also included within a new software solution.

As we mention above, there are many legal changes implemented in the legal software solution within the last few years, but they are just adaptation of existing software solution to the changes of laws that were enacted in the past few years in this area. Some adaptation of the software solution for other services, such as Cadaster service and financial sector within the local government was also made in the past few years.

3 Proposal of a New Information System for Construction Permission Requests as e-service

The analysis of the legal information system of issuing documents for construction permission requests shows some maladjustment of the legal software solution based on the new concept of e-service. For this reason, we made a re-

engineering of the administrative processes in order to improve speed and other performance indicators for the services to the clients. During that process, it is important to be careful with the solution, it has to be web enabled and can be accessed anytime and from anywhere. Also, the solution has to be user-friendly, easy to learn and with a needed help for clients and for the institutional clerks. It has to be safe, secure and to provide enough information for its proper use.

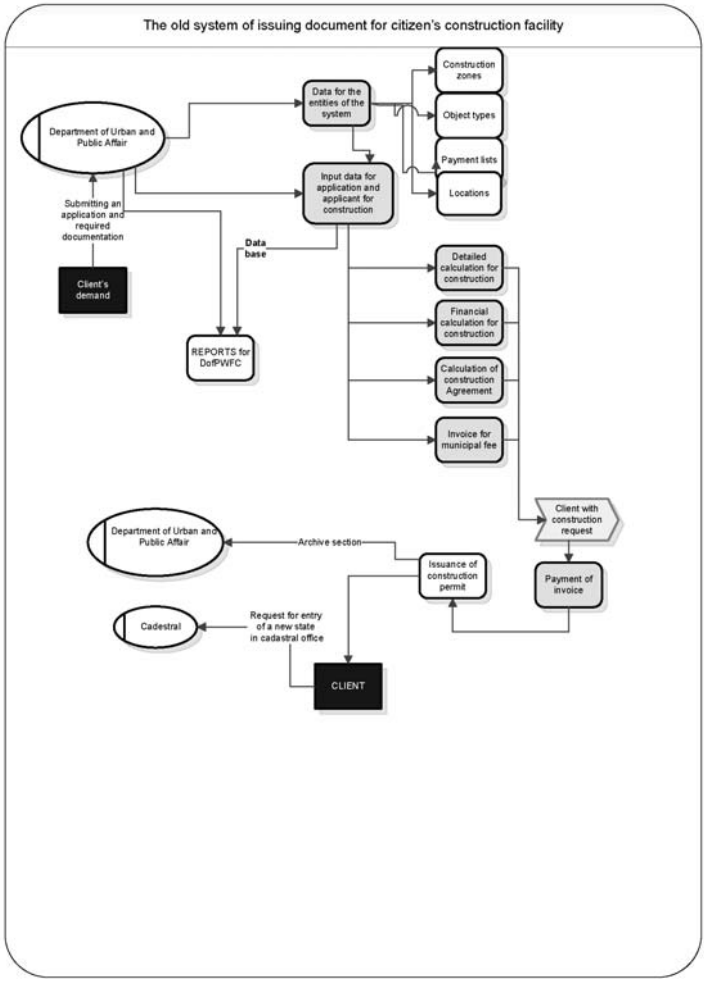

Figure 2 – A legal system for construction permission requests

First of all, it has to be web-based login screen with a possibility of a client registration. For this topic, the web application needs to be created in two official languages (Macedonian and Albanian). After login or registration on the web, the client has to obtain the possibility to enter and edit client's information. Also,

a digital certificate for all users of the system has to be provided because of the system security issues.

After the process of registration, the client must have the possibility to obtain a screen for data input for construction permission request. There must be screens with help files that explain the legal frame, the required documentation and the other facilitators for the proper system operation. In this process, the client has to provide the necessary documents, prescribed in the facility files as PDF or other suitable files for the system.

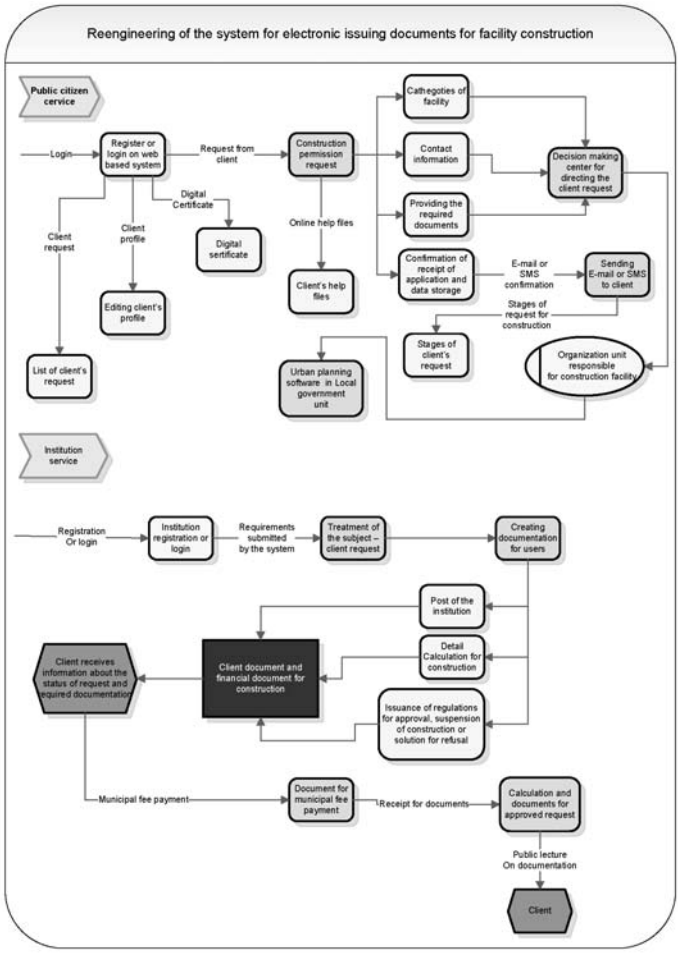

Figure 3 – Reengineering of the Information System for construction permission requests

When the client provides desired data and documents for the system, he has to validate information with a digital signature (certificate), to save the information and to finish the input of construction permission request. After data submission,

the automatic verification for submission has to be provided, sent by e-mail or by SMS. Also, the request stage in database has to be always updated and this information provided for the client.

When the construction permission request is in the database, it has to be send to the decision making center that has to provide in which institution will direct the request. The decision making center has to have its own decision making rules, but the final decision must be done from manager and officer responsible for deciding where the construction permission request will be send. After his confirmation (or alternate decision solution), the name of responsible center have to be connected with the client construction permission request and the legal time for completing the procedure has to start. All changes in the request status have to be placed in database and client must have an insight into the status of his request. The proposed system with the transformed processes as a result of the process of re-engineering is shown on the Figure 3 above. The part of citizen service (clients) is shown on the upper part of the Figure 3.

The second part of the system has to be designed for the institutional service and should be an extension of the previously described processes as public citizen service. The clerk of the respondent institution has to login on (or make a registration on) the system. After the login screen, the assigned request for that institution has to be provided on his screen. The officer has to provide confirmation for acceptance of the request and has to create the desired documents for client construction permission request. The system has to provide clerks of some institutions to get some duties from another clerk for additional information about the client in order to provide some documents (as EVN- Power Company, water and utility local companies etc.). Also, he has to check legal documents, cadaster documents and has to make decision about the approval or refusal of the request.

It is important to mention that the system has to be accessible for many institutions which are the part of the process of confirmation of issued documents needed for the client's request. Institutions that have to have access to the necessary documents for the application such as cadaster, Ministry of internal affairs, private cadaster, EVN, water and utility local companies are some of the institutions responsible for providing documentation on the subject. All of them must have access to the system in order to provide the desired information and documentation.

Finally, all needed documents have to be prepared and send to the client. The urban planning clerk is responsible for preparing the data for final documentation and the documentation must be signed by the clerk and the head of the Department. Also, the invoice for payment of municipal fee has to be provided for the client. After invoice payment and sending the confirmation from the customer, an issuance of the building permission has to be sent to the customer. Also, all documentation has to be send to the clients address in a hard format. These processes are shown on the bottom part of Figure 3.

4 A Need of Maintaining and Updating the Web-based Software Solution

The constant need of maintaining of the software solution is needed for the constant change of the working conditions as well as adaptation of new regulations and amendments to the laws, intervention laws that apply in certain time periods. In this case study we show the adaptation of the Law on the treatment of illegal buildings which was published in Official Gazette No. 23 of 24.02.2011 which goes together with the Law on Urban Planning and Development, published in Official Gazette of RM. No. 51/05 from 30.06.2005. Under this law, the government allowed citizens to legalize all those buildings that were built without obtaining construction permission in the past. The statutory deadline for implementation of the law was to 01.01.2014, but at the request of institutions and citizens is extended to 30.09.2015.

Under this law, administrative procedures adopted by the Government of RM should begin by completing the application for legalization of illegal buildings and should be submitted to the Department of Urban and Public Affairs in the Local Government. The administrative procedures that are provided in the process of legalization of illegally constructed buildings are visually displayed in Figure 4. All the process ends with a Public Announcement of the Awarded Recipients by City Authorities to the client – the owner of the illegal object. In order for this administrative procedure to fit the proposed system, it is necessary to provide the same service in the proposed web-based solution – a new information system. A logical suggestion would be to add new categories of services that are provided when applying for construction permission request. This additional service would only cover services possible with addition of a new service - the legalization of illegally built objects, which will require submission of the same documents as in the case when a client makes a construction permission request through a web -based system. Further processes would be made according to projected pattern of business processes for the issuance of construction permission with dynamic accelerated and shortened deadlines.

The implementation of such a system is certainly not an easy task and requires a commitment of professional software companies that will provide the anticipated tasks and solve the set of emerged problems with web-based software application. To facilitate the operation of such a system which serves as e-service, it should be part of e-government solution that should provide, despite functionality, security assumptions. In order to provide data, encryption web based system should be enabled as well as client and institution's certification. This requirement assumes implementation of encryption with certification authority. If there is a set of the opportunity for a payment method for regular banking online payment or credit cards, it is necessary to introduce additional bank secure payment methods.

5. Real Implementation of Web-based Software Solution for Construction Permission Requests

The project started in 2013 with the support of the Ministry of Transport and Communications and ZELS. The site created for this purpose is already in use, but only some parts of them. It is made by Nextsens, Macedonian IT Company. It is web-based solution for client construction permission requests and other stakeholders in the process (See Figure 5 below). The part of anticipated processes is enabled by this system, but the system is still on the phase of implementation and development. Efficiency of this new web-based system should be improved especially for the clients, with decreasing of the needed time for application and the time to complete the process of submitting a request. It remains to analyze the satisfaction of a client from the new web-based solution (clients as citizens and institutions who use the new system), and satisfaction of other service users of the Department of Urban and Public Affairs, Ministry of Interior, Power company, Water supply and utility local companies in each local government and other entities for which the system is intended (on the side of the institutions). Of course, it has to make benchmarks for its use and efficiency achieved on the part of clients and institutions.

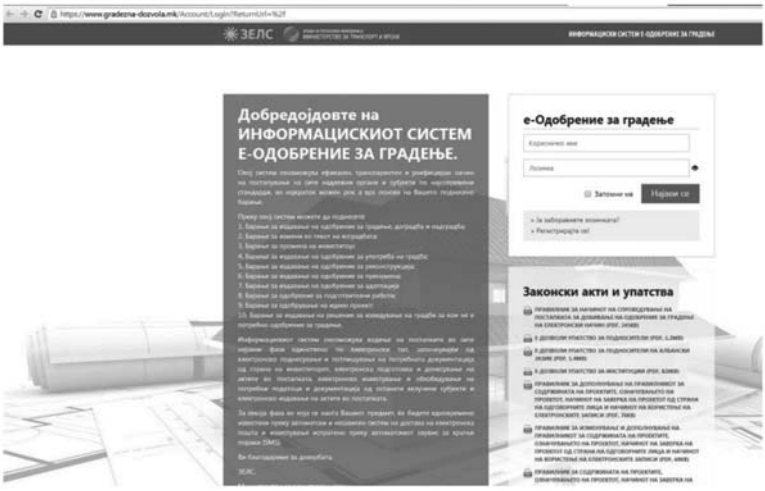

Figure 5 – Screenshot of information system for e-approval for construction

6. E-Service Portal for End-users and Future Improvements

Project e-service for constructiom permission requests is supported by ZELS and the Ministry of Transport and Communications and for this reason it is strongly supported of all state government departments. This project includes

a wide range of institutions involved in the review process and issuance of documents necessary for gaining a construction permission, the holder of the activity (as Ministry of Transport and Communications and the Department of Urban and Public Affairs). While the clents are the service users.

For secure and safe operation of the web system, all stakeholders of the system should possess digital certificates issued by a legal Certification authority. It includes collaboration of clients and institutions with aim of improving public service using web-based technology and providing deadlines for a completion of the process by strictly defined working procedures. Through this solution, the performance of staff responsible for the completion of the process can be monitored as well as the working performance of other institutions.

Using the web based software, some necessary forms and documents for clients are created and send by e-mail in the frame of deadline, followed by sending the documents in a classic way (in the transitional period and completely passed to electronic mode). Only the part of legacy system is not enabled yet with the new software solution. These parts of processes are the responsibility of clerk and are made with word and excel documents.

Collaboration and communication between institutions is also done electronically. In this way, it preserves the system of overall correspondence between institutions providing electronic archive during the actual procedures and processes for each request from the client. These data are very favorable for other activities of the Government of RM which are ongoing and which are supported by the Law of Public Servants adopted on 16.04.2010 and the Law of Administrative Servants adopted on 05.02.2014 and relates to the administration evaluation during the periodical assessment of the staff.

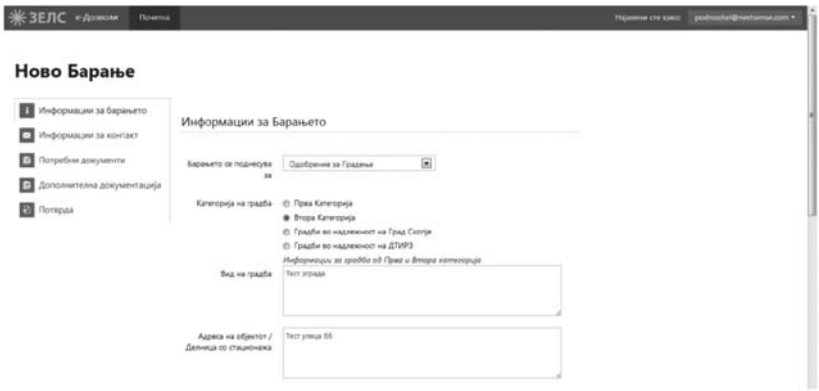

Figure 6 – Screen for Construction permission requests from the web-based system

However, for successful implementation, it is necessary to gain additional and detailed work instructions, good institutional clerk training (they should be included in the system, whether they belong to civil or private institutions as cadaster or other

institutions). However, the most important factor for a successful implementation of the project is the support from the government and its determination to implement this complex project. The project should be related to the whole e-government project as a long term strategy of the Government of RM.

8 Conclusion

From the items explained in the previous text, we can conclude that the actual administrative processes for issuing the Construction permission request in R.M. already are changed. As e part of e-government concept, this web based application connects the clients and institutions involved in the process of issuing permits for construction. These processes have to be improved and upgraded in the future with new services as well as new technology opportunities, such as: mobile application, e-payment and BI tools and software for managers (Ming, 2010). Also, the actual software tools have to be improved with the usage of some visual representation of the data in format suitable for administrative managers in local governments in Macedonia as dashboard or strategic maps of activities.

Besides all, the municipal governments in the Republic of Macedonia must integrate all the needed resources of different nature in order to successfully implement web based software solution that will help the process of issuing construction permits on a municipal level. In addition, there must be a wide spread training of the local public administration servants in different municipal departments of the software use and maintenance.

Figure 7 – Screen for attaching demanded documents for web-enabled system

References

1. Haddad S.& all, Models and Analysis in Distributed Systems, Wiley & ISTE Ltd., 2011
2. Laplante A.P., Ovaska J.S., Real-time System Design and Analysis, Fourth edition, Wiley, 2012
3. McLaughlin B., Pollice G., West D., Head First Object Oriented Analysis and Design, O'Reilly, 2007
4. Ming C. Hao, Daniel A. Keim, Umeshwar Dayal, VisBiz: A Simplified Visualization of Business Operation, HP Laboratory,
5. https://www.gradezna-dozvola.mk/Documents/E-dozvoli_upatsvo%20za%20podnositeli_koregirano.pdf, 24.03.2014
6. http://lokalnirazvoj.rs/assets/files/Baza_znanja/UNDP%20Macedonia%20Local%20Governance%20and%20Decentralization%20Lessons%20learned.pdf, 25.03.2014
7. Law on Administrative Servants, Official Gazette of R.M.,No.27 , adopted on 05.02.2014.
8. Law on Illegal Built Objects, Official Gazette of R.M.,No.23 , adopted on 24.02.2011.
9. Law on Public Servants, Official Gazette of R.M.,No. 52, adopted on 16.04.2010.
10. Law on Urban and Space Planning, Official Gazette of R.M.,No. 51/05 , adopted on 30.06.2005.

EPC – better option for business process generating

Ivaylo Kamenarov, Katalina Grigorova,

Department of Informatics and Information Technologies, University of Ruse,
8 Studentska Str., 7017 Ruse, Bulgaria
ikk@ami.uni-ruse.bg, kgrigorova@ami.uni-ruse.bg

Abstract. This paper presents the similarities between the basic elements of EPC and BPMN. An approach to decompose business processes to individual units and it is recomposing, using process modeling through.

Keywords: Business Process Models, Business Process Management, Business Process Repository, Business Process Generating.

1 Introduction

Nowadays, fast growing business requires from companies to constantly change and adjust their activities and processes appropriate to the business. They need to describe and manage the overall business through business process modeling. Business process modeling should be carried out by a standardized approach to define precise criteria for process description.

Standard EPC (Event-driven Process Chain) allows business process modeling by graphical diagrams, which present the workflow of business processes. Over the years, the standard has improved and established as a powerful tool for modeling, analysis and transformation of business processes and it is used by many organizations. The main elements of the standard are functions, events, connectors and connections between them.

This paper describes an approach for automatic generation of business processes through the description of business processes models with EPC. A short analysis of the similarities and differences between two major and approved standards for business process modeling: EPC and BPMN (Business Process Modeling Notation).

2 Compare EPC and BPMN

EPC and BPMN are the two most used standards for business process modeling by business analysts. Both standards provide means by which it is possible to model almost every business process. Each of which has its supporters. There is a set of rules by which business process model described by a one of standards can be transformed into a model of the same process described by another standard.

The main elements in both standards are largely the same: functions, connectors/gateways, events, control/sequence flow.

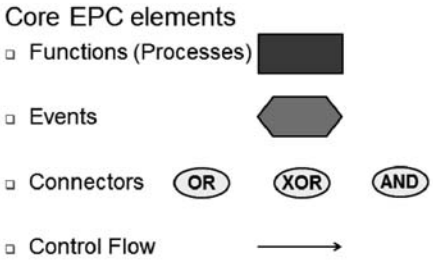

Fig. 1 EPC Elements

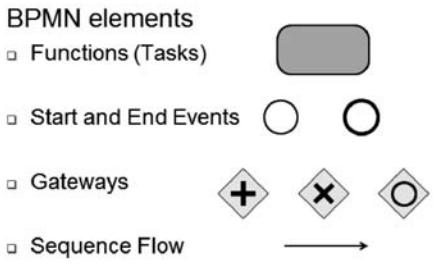

Fig. 2 BPMN Elements

Transformation rules between elements [1].

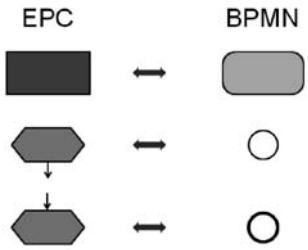

Fig. 3. Rule 1 and Rule 2, respectively, for transformation between functions and triggering / terminating events

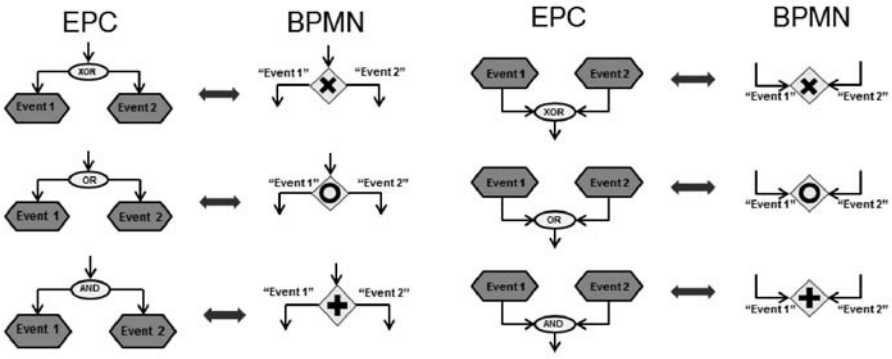

Fig. 4. Rule 3, respectively, for transformation between terminating composite splitting and triggering composite joining events

The Rule 3 (Fig. 4 and Fig. 5) shows that one EPC event is transformed in BPMN sequence flow. There are a semantic difference between EPC events and

BPMN events. EPC events provide a beginning and an end to the process, while BPMN events are: message, timer, link, signal, error etc. If an event in EPC model represents the same sense as BPMN event, then this event is transformed into its corresponding event. For example, if EPC event is a timer, then it is transformed to BPMN timer event.

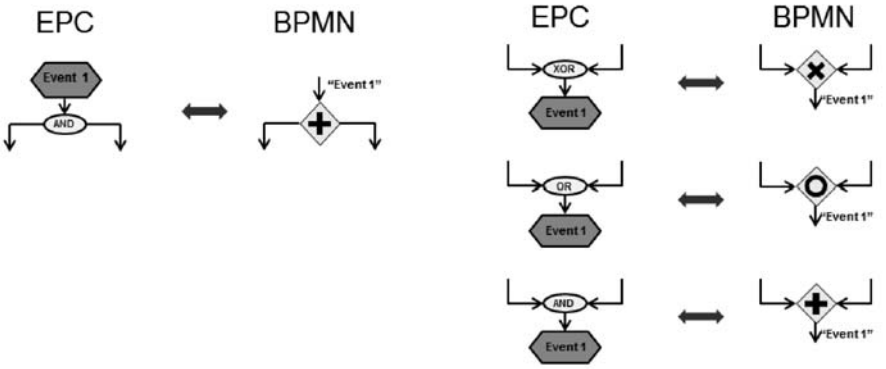

Fig. 5. Rule 3, for transformation between triggering composite splitting and terminating composite joining events

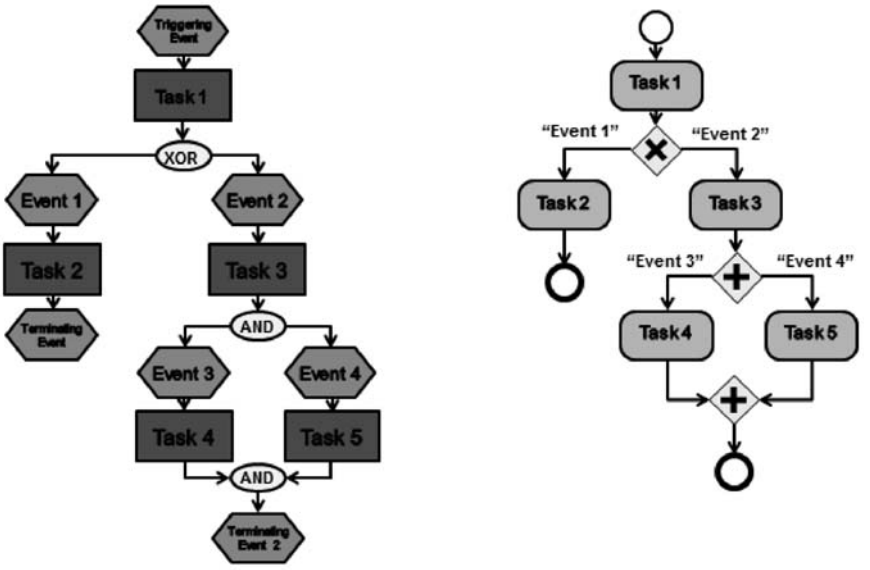

Fig. 6. An example of a business process modeled in EPC and BPMN

Business process generating

In practice is very common to have set of processes, which are arranged in time in some way, i.e. it is not possible to start a process before another is finished. This means that when storing all business processes it is necessary to have a mechanism for presentation and storage of these relationships. Furthermore, together with all the characteristics of the process, it is necessary to store information about the relationships between processes.

The authors choose EPC standard for business process modeling and the elements of created models are stored in a business process repository as elements of EPC.

For each process (function), there is exactly one entry and one exit point [2]. These entry and exit points can be events and connectors, where they are defined respectively as just a simple or a composite event (including connector and events connected to it). In the business processes repository for each process data about its entry and exit points is stored in addition to its characteristics.

A connector groups several related to it events, these events can be both simple and composite. Through connectors, for particular process, it is possible to find all triggering and terminating events. A process has only one entry and one exit point, but there may be more than one triggering events and more than one terminating events.

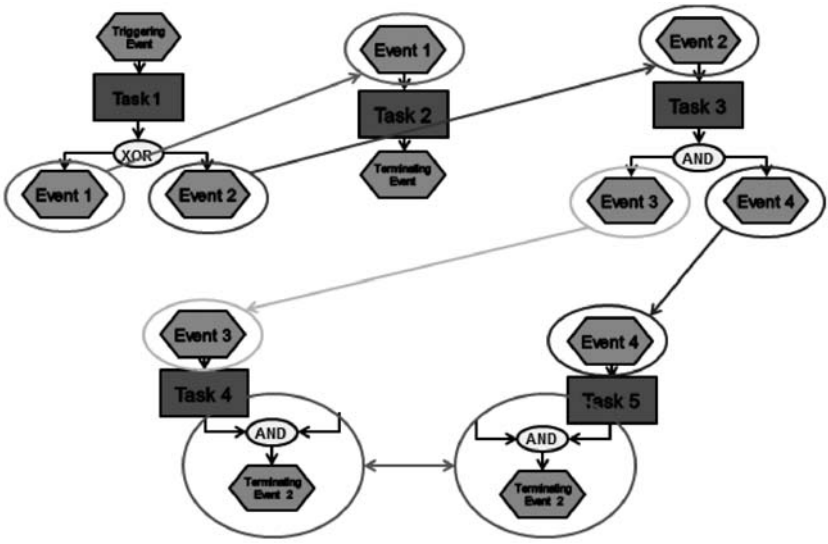

Fig. 7. Process in Fig. 6 with decomposed units

Business processes with all of their triggering and terminating events are complete units that can be used for assembly of a chain (control-flow) of business

process at a higher level. Simple events are used for connecting elements between individual units. An event which appears as a terminating for the process can be used for connecting element to another process that has the same triggering event. An event can be triggering for at most one process and terminating for at most one process. If an event is not terminating for any process (function) then it is a triggering for whole sequence (business process from a higher level), and if an event is not triggering for any process then it is terminating for whole sequence.

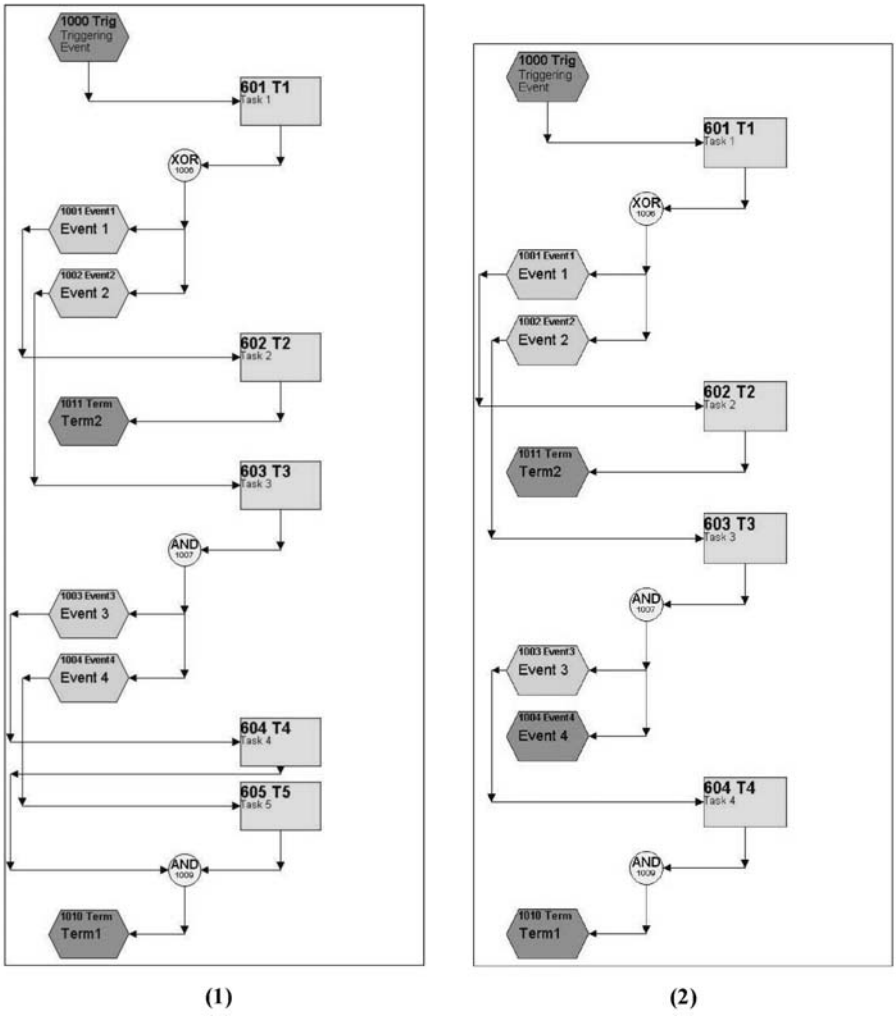

Fig. 8 Example of generated processes

After the decomposition of the business process model, all functions (subprocesses) are associated with triggering and terminating events that represent

complete units. The enclosed event on Fig. 7 is connected with its corresponding event from another unit. So these units predispose for re-compose a process or generate a new one. Thus connections between functions are carried out a natural way by the EPC events.

A prototype of business processes repository has been developed, which provides a possibility for generation of business process through the presented approach. For some reason business analysts need a process model in which a function is not included. In Fig. 8 two example models of the same business process are presented. The difference between the models is that in the second one function "T5" is not presented. This changes the entire model, in the first one there are one triggering and two terminating events, while the second one has one triggering and three terminating events.

The described and implemented functionality offers to analysts a quick and easy way to check and compare different models of a process which includes or not any function.

By BPMN standard [3] it is possible to model the process containing several functions (subprocesses) one after another, without any other components between them, connected only with sequence flow. Thus there is no other option, by which to describe the connections between functions except the sequence flow that connects two components of the model described in BPMN. But sequence flow element is used not only to connect two functions (processes), but also for connecting any elements. This requires introducing an additional mechanism to describe which process should be completed before starting another one. This further complicates the process model and makes its maintenance more difficult.

Conclusion and future work

In EPC for each process there are one entry and one exit points, which may be events or connectors. With them, it is possible to find all triggering and terminating events of the process. These events serve as a connection between the different units.

In BPMN there is no element by which in a natural way to connect pieces of the business process. It is necessary to introduce an additional mechanism by which to clarify the connection between tasks.

When the function is not included in the generated sequence a gap remains in its place. For future work it is planned to develop and implement an algorithm for proper replacement of this function.

The present document has been produced with the financial assistance of the European Social Fund under Operational Programme "Human Resources Development". The contents of this document are the sole responsibility of "Angel Kanchev" University of Ruse and can under no circumstances be regarded as

reflecting the position of the European Union or the Ministry of Education and Science of Republic of Bulgaria.

Project № BG051PO001-3.3.06-0008 "Supporting Academic Development of Scientific Personnel in Engineering and Information Science and Technologies"

References

[1] Willi Tscheschner, Transformation from EPC to BPMN, Hasso-Plattner-Institute, Potsdam, Germany.

[2] W.M.P. van der Aalst. Formalization and Verification of Event-driven Process Chains, Department of Mathematics and Computing Science, Eindhoven University of Technology.

[3] Business Process Model and Notation (BPMN), http://www.omg.org/spec/BPMN/2.0/PDF - May 2014.

Cloud Technologies Application at JINR

Nikita Balashov[1], Alexandr Baranov[1], Nikolay Kutovskiy[1,2], Roman Semenov[1],

[1] Laboratory of Information Technologies, JINR, Dubna, Russia

[2] National Scientific and Educational Centre of Particle and High Energy Physics of the Belarusian State University, Minsk, Belarus
{balashov, baranov, kut, roman}@jinr.ru

Abstract. Cloud technologies are already wide spread among IT industry and start to gain popularity in academic field. There are several fundamental cloud models: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). The article describes the cloud infrastructure deployed at the Laboratory of Information Technologies of the Joint Institute for Nuclear Research (LIT JINR). It explains the goals of the cloud infrastructure creation, specifics of the implementation, its utilization, current work and plans for development.

Keywords: cloud technologies, virtualization.

1 Introduction

The JINR Cloud service was deployed in order to increase an efficiency of the overall IT infrastructure of Laboratory of information technologies functioning: more efficient servers and services management, better hardware utilization, higher services and storage systems reliability. It is build upon an Infrastructure as a Service (IaaS) model. Such model provides network access to computational, software and information resources (networks, servers, storage devices, services and application software), allowing to allocate those resources on-demand according to dynamically changing requirements: cloud users can obtain, configure and deploy cloud services themselves with the minimal assistance of the IT specialists. The Cloud service is expected to reduce the costs of owning the computing infrastructure and also to reduce its support complexity.

2 Service Implementation

The JINR Cloud service is based on an open-source IaaS system OpenNebula [1]. The two main components of the system can be marked out:

- front-end node (FN): contains the system core and user interfaces to interact with the service;

- cluster nodes (CNs): the physical servers which host the users' virtual machines (VMs).

While CNs are the physical machines, the FN is a virtual one hosted on one of the nodes itself.
Two user interfaces are available to access the service:

- command line interface (CLI);
- web-based graphical user interface "Sunstone".

Figure 1 shows the interactions between the cloud service components.

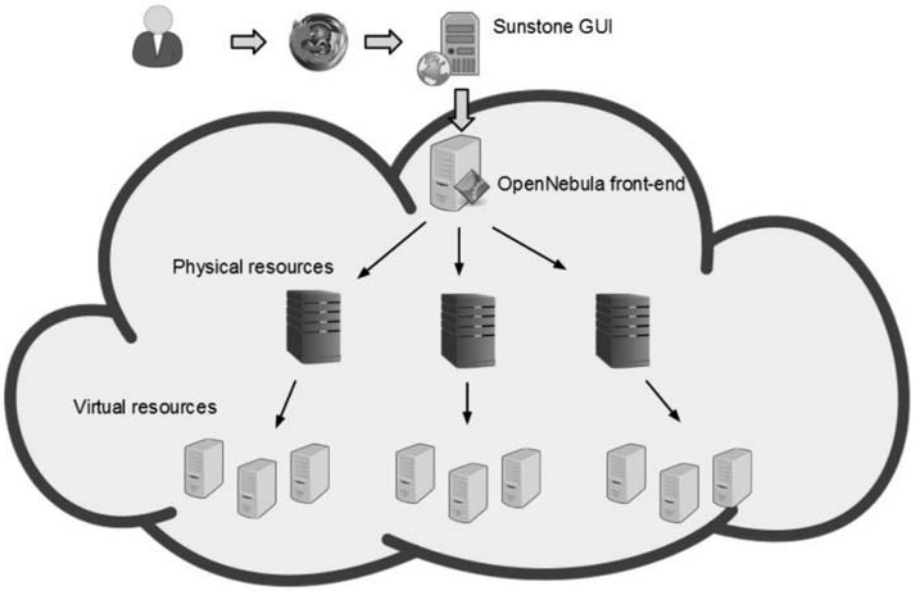

Fig. 1. JINR Cloud service structure scheme showing interactions between its components

Currently the service uses two virtualization technologies to provide VMs:

- OpenVZ [2] (an operating system-level virtualization);
- KVM [3] (provides full hardware virtualization).

The reason why two different virtualization technologies are used is to better fit the variety of the emerging tasks: OpenVZ containers are lightweight and fast but they are bound to use the hosts operating system kernel, while KVM virtual machines support any type of operating systems inside the VMs but have higher overhead.
Originally OpenNebula had no OpenVZ containers support but extensible and modular architecture allowed us to add such support by implementing the custom driver.

JINR cloud service has two types of CNs:

- servers with two mirrored disk drives (RAID1) used to host highly reliable VMs;
- servers with one disk used for educational, research or test VMs.

The functioning of the hosts is monitored by Nagios [4]. Although virtual machines are not monitored currently by Nagios, some their parameters are tracked by built-in OpenNebula monitoring system and its information is available on the Sunstone dashboard.

To make a request on resources or own quotas extension easily for end-users the custom plugin for Sunstone was developed. It's a simple web-form integrated into Sunstone menu. All that web-form's fields need to be filled by the user. Pressing "Send" button automatically generates an email to system administrators for request approval.

The VMs can be accessed either with use of rsa/dsa-key or password. A plugin implementing Kerberos authentication was developed for user authentication in Sunstone. To increase security of the transmissions between the service web-interface and user's browsers SSL encryption is used.

3 Extending OpenNebula

OpenNebula platform is designed to be easily extended with new components and functionality. It has highly modular structure that allows to change the system behavior in many different ways. In particular there is a set of mechanisms allowing to tune the system for your needs. Some of them are the following:

- hooks mechanism that enables triggering of custom scripts on a particular change in a certain resource;
- drivers make possible to add support for new types of resources: storage, virtualization, monitoring, authorization, networking;
- Sunstone plugins make possible to change user interface in any way.

Drivers are just a set of scripts implementing actions defined in OpenNebula API for particular resource.

Sunstone is a Sinatra [5] application having a straightforward implementation making it easy to implement custom plugins.

Using these means JINR cloud service functionality was extended by adding OpenVZ support, Kerberos authentication and quotas request form which were not originally implemented in the OpenNebula platform.

OpenNebula is developing rapidly and some releases, especially major ones, contain changes in data structures processed by the drivers. That leads to the necessity to update custom drivers according to those changes. In order not to maintain backward compatibility of the OpenVZ driver the decision was made to always stick to the latest OpenNebula release. Some OpenNebula releases also contain significant user interface changes and following the latest releases makes user adaption to new interfaces smoother.

4 Service Usage

Currently the service usage is developed in three directions:

- test, educational and research tasks as part of participation in various projects using cloud and grid technologies;
- systems and services deployment with high reliability and availability requirements;
- increasing computing capacities of the grid-infrastructures during peak loads.

The following services and testbeds are currently deployed in JINR cloud:

- EMI-based [6] testbed (used for trainings, performing JINR obligations in international projects such as WLCG [7], etc);
- ATLAS T3MON [8] + PanDA [9] testbed [10] (monitoring tools development for ATLAS Tier-3 sites, PanDA software development for distributed analysis);
- DIRAC-based [11] testbed for BES-III [12] experiment (monitoring tools development for BES-III distributed computing infrastructure);
- DesktopGrid testbed (to estimate the volunteer' computing technology for possible use in solving JINR users' tasks);
- web-service HEPWEB (provides a possibility to use different tools for Monte-Carlo simulation in high-energy physics);
- test instances of the JINR document server (JDS) and JINR Project Management Service (JPMS).

Moreover a set of OpenNebula testbeds are deployed in the JINR cloud service for development and debugging OpenVZ driver for current and new OpenNebula software releases. Each of such testbeds consists of 2-3 KVM VMs:

- one FN of test cloud instance,
- 1-2 CNs with OpenVZ hypervisor installed.

Services and testbeds currently deployed in the JINR cloud are shown in figure 2.

5 Current work and plans

Current work and features to do are listed below:

- implement authentication in VMs through Kerberos;
- create a support mailing list to interact with the end-users (to inform them about news, maintenance, new features, etc);
- estimate the possibility to implement Software as a Service (SaaS) model and/or the ability to provide access to virtual machines with pre-installed applications;
- improve quotas request form;
- deploy web-portal containing HOWTOs, FAQs and other information to improve end-users' experience with JINR cloud service.

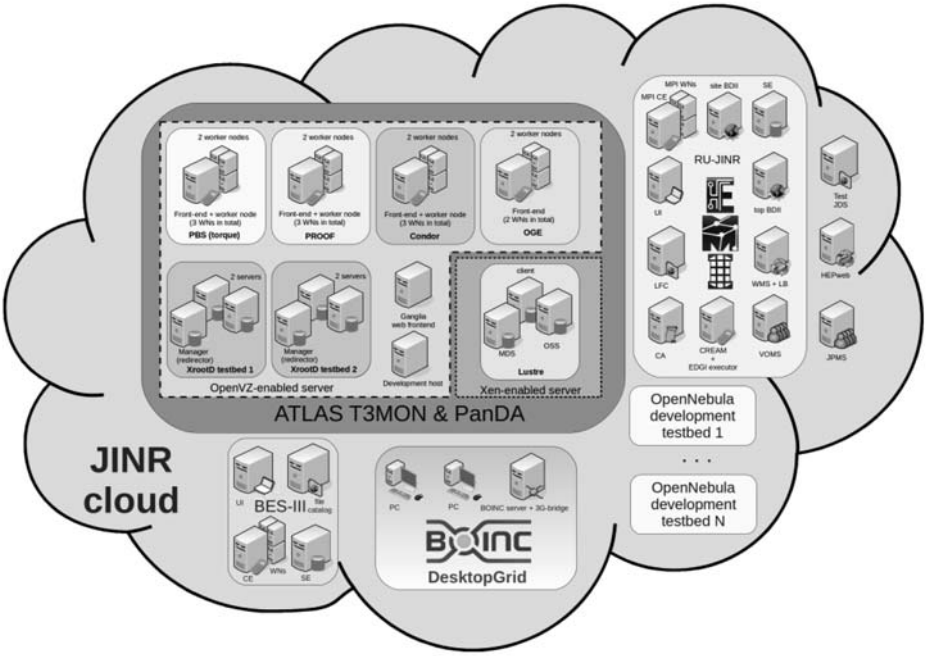

Fig. 2. Services and testbeds currently deployed in the JINR cloud

6 Conclusions

The JINR cloud service made possible to better utilize hardware resources. It also significantly simplified the job of system administrators by automating many virtual machines management tasks and by giving the users the ability to create and manage VMs by themselves within the limit of the granted quotas.

The service is actively used to cover users' demands as well as to carry out JINR commitments in Russian and international projects.

OpenNebula platform showed its stability and the ease of use. The source codes and platform architecture occurred to be well designed and easy to understand that makes it flexible and really easy to extend its functionality with custom plugins and drivers.

References

1. OpenNebula project, http://opennebula.org
2. OpenVZ project, http://openvz.org
3. KVM project, http://www.linux-kvm.org
4. Nagios project, http://www.nagios.org

5. Sinatra project, http://www.sinatrarb.com
6. EMI project, www.eu-emi.eu
7. WLCG project, http://wlcg.web.cern.ch
8. J. Andreeva et al., Tier-3 Monitoring Software Suite (T3MON) proposal, ATLAS note, 2011
9. PanDA project, https://twiki.cern.ch/twiki/bin/view/PanDA/PanDA
10. S. Belov et al VM-based infrastructure for simulating different cluster and storage solutions used on ATLAS Tier-3 sites // Journal of Physics: Conference Series. 2012. Vol. 396. Part 4. 5 pp. doi:10.1088/1742-6596/396/4/042036
11. DIRAC project, http://diracgrid.org
12. Web-portal of BES-III experiment, http://bes3.ihep.ac.cn

Hierarchy and expressions for automated workflows for NGS data processing

Milko Krachunov

Faculty of Mathematics and Informatics, Sofia University, 5 James Bourchier blvd., Sofia 1164, Bulgaria

Abstract. The advent of next-generation sequencing (NGS) technologies has lead to a massive increase in the amount of genomic data available for processing. In some areas of research, such as metagenomics, the pace of data availability has surpassed the pace of development of new computational methods, their testing, compatibility and integration. The latter is a process that is performed by both informaticians and non-informaticians, using both various programming languages and visual tools. This is formulated in the form of workflows that describe the hierarchy of procedures in which every genomic dataset is processed. An automated graphical representation is often sought for visualisation and presentation purposes.

This work presents a YAML-based system and an expression language for describing workflow hierarchies inspired by the Make system for compiling source code. The system is being designed to be susceptible to static analysis which allows automated visualisation and inspection. The use of a standard like YAML enables the development of tools for both automated and computer- aided generation of workflows. At the same time, the system is aimed to be as expressive as possible, without sacrificing its simplicity.

The presented workflow system is a part of larger effort in development and testing of a similarity-based error detection and correction method for metagenomics data, and thus includes extensions to NGS processing algorithms such as fuzzy indicator of reliability, as well as tools such as error simulation procedures.

Key words: NGS data analysis, workflow design, expression language, metagenomics, YAML

1 Parallel sequencing and metagenomics

Parallel sequencing or next-generation sequencing (NGS) technologies deal with the acquisition and processing of significant amounts of genetic sequence data. They offer the means to rapidly determine the order of nucleobases in extensive quantities of DNA fragments from various biological organisms and record them as digital data. The resulting datasets can contain thousands to millions of reads, each consisting of four-letter sequences that can range from tens to hundreds bases in length, and pose a significant challenge during their processing and interpretation, both theoretically and computationally.

The tasks that employ sequencing data can range from the identification of genetic markers in forensics or parental testing, to statistical studies of gene expressions during the research on diseases and their treatments, to search for common mutations in known organisms, to reconstruction of whole genomes from genetic fragments in *de novo* sequencing. [8]

Metagenomic studies in particular are crucial for the ecology and epidemiology. Studies of the microbial communities in soil samples offer important results for agriculture, while studies of such communities inside humans offer important results for the prevention of infectious diseases [7, 10]. Unfortunately, in spite of the high availability of raw data, metagenomic research is set back by the heterogeneous nature of the samples and the lack of established comprehensive procedures, and researchers are often faced with challenges in choosing and integrating various semi-compatible software packages. [12, 14]

The aim of this paper is to present a solution for integrating different processing tools into a flexible workflow system. Such system would greatly facilitate many researchers in bioinformatics, particularly those dealing with metagenomic data.

V. Dimitrov, V. Georgiev (Editors): ISGT'2014. ISSN 1314-4855
Proceedings of the 8th International Conference on
Information Systems and Grid Technologies, Sofia, 30–31. May, 2014.

2 Workflow descriptions

The aim of this work is to present a solution for describing hierarchy of operations that are found in parallel sequencing workflows applied in metagenomics. They provide the means to create the necessary software glue between arbitrary processing programs in the form of a parametrised workflow. The workflow descriptions are descriptive, susceptible to parallel and distributed execution, as well static analysis—both by automated inspection and automated visualisation—and allowing complex processing applications like shown on Figure 1.

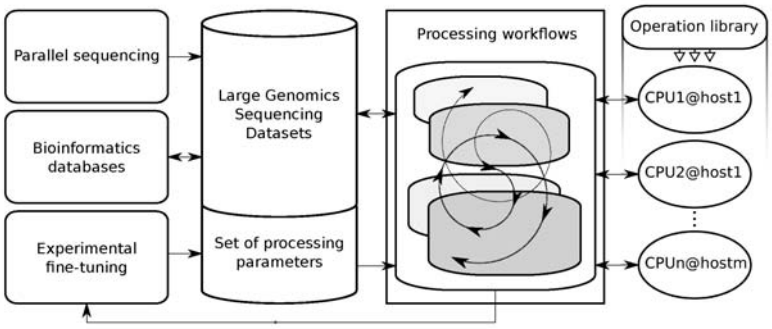

Fig. 1. Application of a complex workflow

The system is inspired by the Make system [13] for compiling software projects, which is popular among bioinformaticians. The processing is split into targets that need to be produced, and that provides an acceptable level of simplicity in the descriptions, while also making distributed and parallel execution easy. The description format uses the YAML object notation language, [1] which offers a standard way for the workflows to be read or generated with the existing tools that are available for the majority of programming languages, ensuring easy interoperability.

A terse C-based expression language that is interchangeable with a verbose YAML description is also being introduced for describing the conditional dependencies inside a switched workflow with sacrificing the clarity of the datasets. To maintain the compatibility with external tools, the expressions can be converted into the verbose YAML objects, and their syntax is kept simple to enable the generation of such expressions using external tools.

Using a whitelist of allowed operations that are confirmed to be safe, safe execution of unverified workflows from untrusted sources is made possible.

2.1 Target datasets definitions

Each workflow description consists of a collection of target dataset definitions. These definitions can refer to input datasets that are required before the execution of the workflow and are provided by the user, or they can apply a processing operation over a group of other datasets to construct the target dataset. The operations are taken from a predefined set that can be extended with user-defined operations which asynchronously execute arbitrary tools for genomic data analysis. Each operation definition offers the necessary descriptiveness to allow for implicit file format conversion to take place when required.

The application of operations can be simple—applied over a single dataset, mapping—applied over a collection of datasets to produce a processed collection of datasets, or reducing—applied over a collection of datasets to merge them into a single one.

Each value, dataset, or dataset target is defined as an element inside the workflow's YAML associative array. Each element is also an associative array describing the value or dataset, and the '`category`' key is common for all descriptions, and defines whether the element refers to an input, a simple, mapping or reducing operation. The description elements can use one of the following signatures—using the '`input`', '`operation`', '`map-operation`' or '`reduce-operation`' categories respectively.

Inputs. An *input value* or *input dataset* need to be explicitly available before the execution of the workflow, either in a file or internet URL specified by the user, or as a result during the execution of another workflow. The inputs are considered to be the arguments and parameters of the workflow.

```
source: {category: input, optional: false, type: biosequences, key: source}
```

This example defines the '`source`' input dataset inside the workflow. By default, the parameter name used to pass it to the workflow is the same at the dataset name inside the workflow, but the former can be overridden using '`key`'.

To specify default input values, the '`optional`' key needs to be '`true`', and a default value needs to be passed under the '`default`' key. Complex input datasets can also be made optional, but they cannot have an inline default value—they need to be either '`null`' (the default), or a reference to another dataset, using the '`!ref`' custom YAML type (see 2.1).

The input is implicitly defined to have a type of '`BioSequences`'[1], which is needed to signify how the input will be read from a file.

Targets produced by a simple operation execution. A value or dataset that is defined with the *operation* category is implicitly calculated with a simple operation execution. While, similarly to the Make compilation system, implicit datasets can be cached, unlike it, no explicit values can be provided for them by the user before the workflow is executed.

```
aligned:
  category: operation
  method: alignment-multiple:align
  arguments:
    sequences: !ref source
    quality: 0
```

```
clustered:
  category: operation
  method: clustering:cluster
  arguments:
    sequences: !ref aligned,
    threshold: !optional-input threshold
```

These two examples define two implicit datasets. The '`aligned`' target dataset is constructed using the '`align`' method of the a '`alignment-multiple`' service, which selects any [2] suitable multiple sequence aligner available on the local machine, and in case of a distributed executor—on some remote machine that is presently idle. The method is called with two arguments, with a reference to the '`source`' dataset from 2.1 being passed under the '`sequences`' one, using the '`!ref`' custom YAML type discussed in 2.1. The '`clustered`' target dataset is then constructed from '`aligned`' one after the execution of a sequence clustering service. Dataset types do not need to be specified, if they are needed for inspection purposes, the operation descriptions from 4.2 can be used to obtain them.

[1] The types in the type hierarchy and their implementations use CamelCase names, but they are referenced in the workflows using lowercase aliases.

[2] In a non-example workflow a service that would select a specific aligner would be used, because the choice of aligners is very important for the results of the study. This is discussed in section 4.2.

Targets produced by a mapping and reducing operations. A dataset that is defined under the 'map-operation' category is a sequence of items (datasets or values) implicitly calculated as a result of applying a simple unary operation over another sequence of items in which a collection of objects is processed into another collection of objects. Similarly, a 'reduce-operation' merges a collection of objects into a single one using a binary operation applied to the pairs in the collection.

```
realigned:
    category: map-operation
    iterate-over: !ref clustered
    method: alignment-multiple:align
    arguments: {sequences: !current}
```

```
merged:
    category: reduce-operation
    iterate-over: !ref realigned
    method: alignment-group:align
    arguments: {left: !left,
                right: !right}
```

The first example defines a 'realigned' dataset which aligns the clusters generated in 2.1 one by one using the 'alignment-multiple' provider, and each cluster is passed as the 'sequences' argument of the 'align' method. The second example defines a 'merged' dataset which merges the clusters from 'realigned' using a provider of the group alignment service, merging pairs of clusters into one until there's only one dataset.

A special YAML type '!current' references the current element in the iteration. Its value can be any empty scalar value, such as 'null', or '', and whenever YAML syntax allows it - nothing at all. Non-empty and non-zero values shouldn't be used, as they are reserved for potential extensions that would allow nested mapping operations using values to reference the current value in each respective loop.

As reduction requires binary operations, the current elements are references using two special YAML types—'!left' and '!right' that work the same way as '!current'.

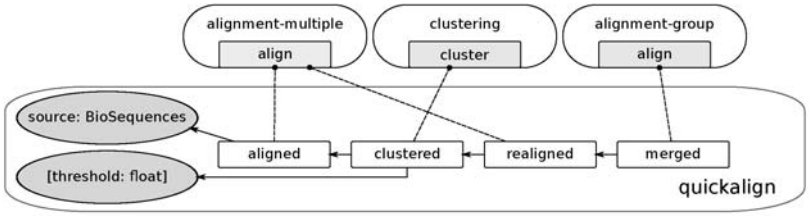

Fig. 2. Example"quickalign" workflow

Combining all the descriptions from the examples above would result in a workflow describing the common common procedure used in Metagenomics to quickly align extremely large datasets that are not susceptible to direct alignment at the expense of a moderate decrease in the quality. A visual representation of that workflow is shown on Figure 2.

References to datasets. As shown on the examples, values and datasets can be references within the workflows using the custom YAML type '!ref'. To facilitate the implementation of workflow execution and inspection tools, these references are only allowed as arguments to methods and, as an exception, as default values for input datasets. Potential extensions include references included inside a composite datatypes such as associative arrays (for the use in switch-case operators), as well as passing a reference as service name when specifying a method (for parametrised service selection), but these are prohibited in the current version.

There are two more ways to reference other datasets. The '`!expr`' extension declares an expression and references its result as discussed in 3, and the '`!input`' and '`!optional-input`' extensions declare a new input in-place and reference it—it can be declared using either its key (a string), or a full associative array as in in 2.1.

Additionally, any optional parameters of the underlying operations are automatically exposed as in-place inputs. For example, if the operation constructing the '`clustered`' target accepts an unspecified '`threshold`' argument, it is exposed as the '`clustered.threshold`' input.

3 Expression language

3.1 Expression syntax

For expressing complex conditional processing in workflow branching, an expression language is defined. It is defined by a context-free grammar that is compiled by an LR(1) parser. The grammar resembles C expressions, using the standard C operators extended with a power operator, but using an extended alphabet for identifiers. Using extended[3] Backus-Naur Form it can be written as following.

⟨*integer*⟩ ::= ⟨*digit*⟩ { ⟨*digit*⟩ } ; ⟨*float*⟩ ::= ⟨*digit*⟩ { ⟨*digit*⟩ } . { ⟨*digit*⟩ } ;
⟨*identifier*⟩ ::= (⟨*letter*⟩ | _) ⟨*letter*⟩ | ⟨*digit*⟩ | _ | : | - ;
⟨*dataset*⟩ ::= ⟨*identifier*⟩ ; ⟨*function call*⟩ ::= ⟨*identifier*⟩ ([⟨*expr*⟩ { , ⟨*expr*⟩ }]) ;

⟨*unary*⟩ ::= (~ | - | +) ⟨*expr*⟩

⟨*multiplicative*⟩ ::= ⟨*expr*⟩ (* | / | // | %) ⟨*expr*⟩ ; ⟨*additive*⟩ ::= ⟨*expr*⟩ (+ | -) ⟨*expr*⟩ ; ⟨*power*⟩ = ⟨*expr*⟩ ** ⟨*expr*⟩ ;

⟨*bitwise and*⟩ ::= ⟨*expr*⟩ & ⟨*expr*⟩ ; ⟨*bitwise xor*⟩ ::= ⟨*expr*⟩ ^ ⟨*expr*⟩ ; ⟨*bitwise or*⟩ ::= ⟨*expr*⟩ | ⟨*expr*⟩ ;

⟨*shifts*⟩ ::= ⟨*expr*⟩ (>> | <<) ⟨*expr*⟩ ; ⟨*comparisons*⟩ ::= ⟨*expr*⟩ (< | <= | > | >= | == || !=) ⟨*expr*⟩ ;

⟨*logical not*⟩ ::= ! ⟨*expr*⟩ ; ⟨*logical and*⟩ ::= ⟨*expr*⟩ && ⟨*expr*⟩ ; ⟨*logical or*⟩ ::= ⟨*expr*⟩ || ⟨*expr*⟩ ;
⟨*conditional*⟩ ::= ⟨*expr*⟩ ? ⟨*expr*⟩ : ⟨*expr*⟩ ;

⟨*expr*⟩ ::= ⟨*atom*⟩ | ⟨*power*⟩ | ⟨*unary*⟩ | ⟨*multiplicative*⟩ | ⟨*additive*⟩ | ⟨*shifts*⟩ | ⟨*bitwise and*⟩ | ⟨*bitwise xor*⟩ | ⟨*bitwise or*⟩ |
⟨*comparisons*⟩ | ⟨*logical not*⟩ | ⟨*logical and*⟩ | ⟨*logical or*⟩ | ⟨*conditional*⟩ ;
⟨*atom*⟩ ::= ⟨*dataset*⟩ | ⟨*integer*⟩ | ⟨*float*⟩ | ⟨*group*⟩ | ⟨*function call*⟩ ; ⟨*group*⟩ ::= (⟨*expr*⟩) ;

Priority of operators and other syntax elements such as their order of associativity are handled by the used LR(1) parser, lrparsing for Python. In particular, the priority is as the order in ⟨*expr*⟩, and the power operator, all unary operators, bit shift and bitwise operators are right-associative. All operators within the same group have the same priority.

3.2 Expression usage

Inside a workflow description, expressions can be used as operation arguments in place of dataset references using the '`!expr`' extension. When a workflow is loaded, each expression is compiled into a set of workflow dataset descriptions, they are inserted into the dataset hierarchy, and they are evaluated in the same manner as the corresponding workflow would.

Each operator has its own method within a special '`core`' service, which is called with the specified operands during the evaluation. For basic value types like integers, these methods apply the standard operator. For datasets and other datatypes, Python operator overloading allows

[3] Using brackets for optional elements, braces for zero or more repetitions and parentheses for grouping, and semicolon for ending a rule.

the execution of common operations between datasets. These operators can asynchronously launch long processing procedures such as dataset merging, but to avoid confusion none such overloads are defined for the standard datasets. Function calls can reference any service method, and are preferred for long asynchronous operations.

Unlike most service methods available to workflows that depend on all the datasets passed as their arguments on execution, the arguments of the logical operators and the conditional operator are always lazily evaluated when they become needed and can be used for creating workflow branches. For example, the following conditional expression can be used to pick a quick processing for large datasets and quality processing for small datasets. If '`dataset`' has over 3000 elements, only the '`processed_quick`' dataset is constructed, and if it is below 3000 elements, only '`processed_quality`' is constructed.

```
counting:count(dataset) >= 3000 ? processed_quick : processed_quality
```

The compilation of the expression produces the following workflow dataset descriptions that are added to the hierarchy, and the expression becomes a reference to the last one. The second and third argument of '`core:condition`' are lazily evaluated on demand.

```
_count_dataset_x002: {
  arguments: [!ref dataset],
  category: operation
  method: counting:count

_count_data_ge_3000_x001:
  arguments:
    [!ref _count_dataset_x002, 3000]
  category: operation
  method: core:ge

_if_count_data_processed__processed__x000: {
  arguments: [!ref '_count_data_ge_3000_x001',
              !ref 'processed_quick', !ref 'processed_quality'],
  category: operation, method: core:condition }
```

4 Operations and data types

The workflow executor implementation has to offer a collection of operations that can be used within the executed workflows. While some operations can be implemented inside the executor itself, to ensure wide applicability and extensibility it is necessary to provide the ability to declare arbitrary operations through the use of external tools. For that reason, a YAML format for declaring operations is defined.

The available operations consist of a prototype with a call signature which is exposed to the workflows on one hand, and an implementation or implementation glue on the other. There is also a typing system that is used for the prototypes. Because the typing and call signatures are not crucial for the execution and inspection, they are discussed only briefly here and aren't formally defined.

4.1 Data type descriptions

The descriptions of the data types allowed within workflows are built on top of the Python class system [2]. In Python each class A can have multiple direct ancestors [4] and descendants. In addition to that, abstract Python classes [11] allow class membership to be determined by arbitrary expressions, and Python metaclasses allows operator overloading to be applied over dynamically-generated classes. Using that, defining class arithmetic over a collection of hollow classes without implementation for the purposes of defining method prototypes is straightforward.

[4] Multiple inheritance is supported

Let's denote a subclass relation with $A \subset B$, and class membership relation with $x \in A$. If a hollow class no implementation, it is solely defined by its parent classes, so the entire definition of a class A that inherits classes B_1, ..., B_n can be summarily written as following.

$$A : A \subset B_1, A \subset B_2, \ldots, A \subset B_n \tag{1}$$

Since the basis of our class hierarchy does possess the expected features of sets when it comes to transitivity and membership, it is also true that $\exists i : x \in B_i \Rightarrow x \in A$, and that $X_1 \subset X_2 \wedge X_2 \subset X_3 \Rightarrow X_1 \subset X_3$.

Overloading the subclass relation and the needed operators between classes, we define the following extensions.

Class itemization Overloading itemization, we define the dynamic class '`A[B]`'. $A[B]$ refers to containers of class A that contain objects of class B. Object x is a member of $A[B]$ if and only if x is a member of A and $x[k]$ is a member of B for every possible k.

$$x \in A[B] \Leftrightarrow x \in A \wedge \forall i(x_i \in B) \tag{2}$$

$$A'[B'] \subset A[B] \Leftrightarrow A' \subset A \wedge B' \subset B \tag{3}$$

Thus defined, itemization is an associative operation, and $A[B[C]]$ is equivalent to $A[B][C]$.

Set operations Overloading the bitwise operators that are commonly used for bitsets and regular finite sets in various languages, we introduce the dynamic classes '`A | B`', '`A & B`' and '` A`' as union, intersection and complement of classes.

4.2 Method prototypes and services

The thus defined type extensions provide a language that is rich enough for the operation method prototypes that are in use within the developed workflow system.

The operation methods are grouped into services. Services are similar to interfaces in languages like Java. A service is a collection of method signatures, and each service can have multiple implementations called providers—like an implementation on the local machine, and an implementation on a remote machine in a cluster. Services are registered in the workflow system under a given name. The providers can be registered with the services dynamically—for example, in a distributed executor when a new machine connects to the cluster.

The service and method name are usually separated by a colon, and allow the following characters.

⟨*service*⟩ ::= ⟨*letter*⟩ ⟨*letter*⟩ | ⟨*digit*⟩ | - ; ⟨*method*⟩ ::= (⟨*letter*⟩ | _) ⟨*letter*⟩ | ⟨*digit*⟩ | _ ;

As an example, let's take the '`alignment-multiple`' service that was used in the workflow operation from 2.1. It exposes a single method '`align`' that takes a collection of sequences as an argument. Assuming that the sequences can be passed either as a sequence file referred to by a '`BioSequences`' object, or a '`Sequence`'[5] of '`BioSequence`' objects, and always returns a '`BioSequences`' object, the method would have the following signatures.

```
alignment-multiple:align(sequences:
                         BioSequences | Sequence[BioSequence]): BioSequences
```

[5] Any Python sequence of objects in the default implementation.

Alternatively, a hollow class 'BioSequencesLike' can be defined with the given expression and used in the signature. One potential extension of the typing system is to allow the definition of convertors. As an example, a convertor turning 'Sequence[BioSequence]' and 'Sequence[string]' into 'BioSequences' can be defined, which would make the conversion of the types automatic for methods with only 'BioSequences' in the signature.

4.3 Definition of custom services

There are several ways to define new services or providers. When a workflow is loaded, it is automatically registered as a service and a provider for it. At the same time, using YAML descriptions, command line utilities can be registered as new services or providers into the workflow system.

External services using templates. The standard way to define a new service is using a template for declaring operations of a common type, for example multiple alignment. The template has its own parent service that refers to all services registered under using this template, and each time the template is used, a new service and provider are used.

```
alignment-multiple-mafft:
  factory_name: alignment-multiple-cmdline
  output_on_stdout: true
  commandline: mafft {options} {input}
  options: {nofft: !switch --nofft,
            ep: !option:int --ep {0:d}}
```

```
alignment-group-mafft:
  factory_name: alignment-group-cmdline
  output_on_stdout: true
  commandline: >
    mafft-profile {left} {right}
  options: {}
```

These examples defines two new services using the MAFFT [4] aligner. The first example, using the 'alignment-multiple-cmdline' template, creates a new sub-service under 'alignment-multiple' called 'alignment-multiple-mafft'. When the align method is called, the 'mafft' executable is called with the sequences in written to the 'input' file in the FASTA format, and the result is read from the standard output.

Also two optional arguments for the method are defined. The 'nofft' accepts a boolean value, which when true adds the '--nofft' switch on the command line. The 'ep' argument accepts an integer which is passed under the '--ep' command line option.

Similarly, the second example defines a group alignment that has a binary method accepting two sets of sequences as 'left' and 'right' argument respectively.

Defining new templates. Services and providers can be added using the template factories, however to create a truly custom service, the workflow system allows the creation of arbitrary new templates. The following example shows how you can define the group alignment template used above.

```
alignment-group-cmdline: {
    service_name: alignment-group, operation_method_name: align,
    input_types: {left: biosequenceslike, right: biosequenceslike}
    input_convert: {left: biosequences, right: biosequences}
    output_types: {output: biosequences}
    default_input_formats: {left: fasta, right: fasta}}
```

This defines a new template, having a binary base service 'alignment-group' and a factory 'alignment-group-cmdline'. The service and its subservices offer a single binary method 'align' taking two arguments—'left' and 'right' which are substituted into the corresponding placeholders within the command line. The accepted and final input types are specified, as well as the output type, together with the file format they are encoded in.

Workflow services. Workflows and services are interchangeable. Every time a workflow is loaded into the system, it is automatically registered as a service and its own provider. Every dataset of the workflow can be generated by calling a method of the same name, and every input to the workflow can be passed as an argument. Typically, only one of the datasets needs to be flagged as the main dataset and used as a method in the service, but the workflow system places no such limitation.

This not only allows the creation of complex services that constitute an entire workflow, but it also allows the creation of recursive workflows. Each workflow is allowed to execute itself. That can be used, for example, to improve the cluster align workflow from Figure 2 whose descriptionswere given above starting in 2.1. Instead of aligning the clusters with an external aligner, the workflow can align them with itself, splitting each cluster into further clusters until some condition has been reached.

5 Implementation

The workflow system has a reference implementation written in Python, using the Twisted [15] framework which facilitates asynchronous execution of the external tools and provides the networking functionality for the planned distributed computing extensions. Once the specification and implementation are mature enough, it will be released as free software under the X11 license[6].

The staging setup can perform basic metagenomics data preparation. It is used to perform multiple alignment using MAFFT [4] and MUSCLE [3], sequence clustering through CD-HIT [9], group alignment using MAFFT, error detection and correction using an algorithm developed in-house [5, 6], and can also be extended to provide provide these with arbitrary external command line tools. These tools are combined into a workflow that provides preprocessed and denoised data for carrying out metagenomic research.

5.1 Workflow dependency resolution and execution

The workflow can be described using a dependency graph. Each dataset will have dependency edges directed towards the datasets required during the operation constructing it. For processing operations these would all be hard dependencies, and for branching operations and expressions these would be conditional dependencies that would only be required depending on a conditional test.

For visualisation purposes, a full graph with both types of edges would need to be constructed to provide a full picture of the workflow, with different colours for the different types of edges. The full graph is also useful for validating the data types across the processing chain. In a full graph without any conditional dependencies it is also possible to validate against cyclic dependencies that would lead to infinite loops. This is not possible in the presence of conditional dependencies because of the unsolvability of the halting problem, however ignoring graph loops that can be cut at branching dependencies provides is one way to provide partial validation.

During the execution in the reference implementation, the dependency graph is constructed only with the hard dependencies to ensure any errors—such as cyclic dependencies—are caught early on, and any conditional dependencies are added to it as required. The implementation is

[6] http://www.xfree86.org/3.3.6/COPYRIGHT2.html#3

also written to support future dependency resolution extensions where necessary, for example the definition and use of automated data type convertors where necessary.

5.2 Application of the workflow descriptions

The presented system is expressive and flexible, extensible and reasonably simple. Arbitrary operations can be defined, which allows the use of wide variety of tools inside the workflows. The provided expression language and the respective core operations provide flexibility through complex branching rules with high degree of expressiveness. The description language is split into punctual target-generation call blocks that are easy to define and read, while the use of a standard language like YAML makes it straightforward to generate them by semi-automated and graphical means.

It is not difficult to implement software tools supporting the workflow format to read, write and edit workflows. The reference implementation can be used to remove or annotate any expressions, reducing the format to basic YAML that can be universally understood by any YAML library without the need to parse expressions. Since expressions are only used as branching conditions, their contents are not necessarily significant for manipulating the workflow.

Anything in the workflow can be configured through parameters. A workflow author can make any anticipated settings exposed an explicit parameter, and any functional parameters are exposed as parameters implicitly. There are straightforward and forward-compatible ways to introduce even more implicit parameters—like preferred alignment service—making the system even more configurable.

Workflows can easily and securely be shared between users by maintaining a set of verified operations that are allowed within a workflow. While any additionally defined operations need to be verified to ensure the secure execution is maintained, they are defined in concise execution blocks that facilitate verification, and all parameters are passed securely to avoid any unexpected security issues.

Workflows can also reference one another, allowing their combination into larger workflows. They provide most of the features required for implementing arbitrary workflows in the processing of parallel sequencing data, as well as other tasks in Bioinformatics.

6 Conclusion

A format for manageable workflows has been defined for the use in parallel sequencing data analysis. The format aims to match the requirements for expressivity, flexibility and simplicity that are present in such studies. A new expression language has been formally defined to declare workflow branches. Standardised ways to define new operations within workflows have been included. A reference software implementation has been developed, and it is already used for the preprocessing of data for metagenomics research.

The presented system can be applied as integration tool in metagenomic sequencing data analysis, facilitating the use of various combinations of processing tools, their experimental validation and their sharing in a standard way.

Acknowledgements

The presented work has been partially funded by the Sofia University SRF within the "New approaches and information technologies for building special-purpose knowledge based systems" Project, Contract No. 044/2014.

References

[1] Ben-Kiki, O., Evans, C., döt Net, I.: YAML Ain't Markup Language (YAML) version 1.2, 3rd edition (Oct 2009), http://yaml.org/spec/1.2/spec.html, [Online, accessed 18 December 2013]

[2] Chaturvedi, S.: Python types and objects (2009), http://www.cafepy.com/article/python_types_and_objects/index.html, [Online; accessed 27 Dec 2013]

[3] Edgar, R.: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5(1), 113 (Aug 2004)

[4] Katoh, K., Kuma, K., Toh, H., Miyata, T.: MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleid Acid Research 33(2), 511–518 (2005)

[5] Krachunov, M.: Denoising of metagenomic data from high-throughput sequencing. In: Advanced Research in Mathematics and Computer Science. pp. 67–76. Sofia (2013)

[6] Krachunov, M., Vassilev, D.: An approach to a metagenomic data processing workflow. Journal of Computational Science 5, 357–362 (2014)

[7] Kristensen, D., Mushegian, A., Dolja, V., Koonin, E.: New dimensions of the virus world discovered through metagenomics. Trends in Microbiology 18(1), 11–19 (Jan 2010)

[8] Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., Wang, J.: De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20(2), 265–272 (Feb 2010)

[9] Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13), 1658–1659 (Jul 2006)

[10] Nelson, K., White, B.: Metagenomics and its applications to the study of the human microbiome. Metagenomics: Theory, Methods and Applications pp. 171–182 (2010)

[11] Guido van Rossum, T.: Introducing abstract base classes (Jan 2009), http://www.python.org/dev/peps/pep-3119, [PEP 3119]

[12] Scholz, M.B., Lo, C.C., Chain, P.S.G.: Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Current Opinion in Biotechnology 23, 9–15 (2012)

[13] The Austin Group: make, The Open Group Base Specifications Issue 6, IEEE Std 1003.1. The IEEE and The Open Group, 2004 edn. (2004), http://pubs.opengroup.org/onlinepubs/009695399/utilities/make.html

[14] Valverde, J., Mellado, R.: Analysis of metagenomic data containing high biodiversity levels. PLoS ONE 8(3) (2013)

[15] Zadka, M., Lefkowitz, G.: The twisted network framework. 10th International Python Conference (2002), https://twistedmatrix.com/users/glyph/ipc10/paper.html

Re Pub(lic) of Philo(sophy)

Mícheál Mac an Airchinnigh | Михал Мак ан Аирхини

<mmaa@cs.tcd.ie>
School of Computer Science and Statistics
University of Dublin, Trinity College
Dublin 2, Ireland

Vassil Nikolov | Васил Николов

<vnikolov@pobox.com>

Abstract. One gets the impression that our modern electronic world is strictly and intrinsically impersonal. There is nothing new in this impression. It applies equally well to older technologies, such as radio, telephone, and television. All the Information that is available to us may be roughly classified as those of the 6 senses: aural (ears to hear), visual (eyes to see), tactile (hands to touch) (such as use of brail for the blind), smell (nose to smell), taste (tongue to speak and to taste), ... To these physical 5 we add a sixth sense. Call it what you will: imagination/intuition/… Sense is made of our world through the mediation and interpretation of the central nervous system in the brain. Traditionally we also speak of the 6 ages of "Man," the latter being a generic term for "human." Shakespeare covered these well in his play "..." ?Much ado about nothing? Today in 2014, we are fascinated by Big Data, the cocoon of the Cloud, "always on" internet, "electronic friends," unlimited access to information of all kinds and in all languages, natural and otherwise. We also recall two ancients who have been instrumental in the establishment of current Western Culture: Plato and Aristotle. These two embody the two ways of … Theory and Practice.

Keywords. Collection, Dissemination, Ontology, Processing, Sense, Storage

1. Introduction

At the beginning of the Age of Computing [AC], as we know it, one had to know how to encode "well thought out" algorithms. These encoded algorithms were then processed on Analogue Computers [AC] (Wikipedia Editors, 2014c) and Digital Computers [DC] (Wikipedia Editors, 2014e), hopefully to execute without error. Validation of the execution was a *sine qua non*. One deliberately plays on the intrinsic ambiguity of [AC] and [DC], in a humanistic cultural fashion as counterpoint to the perceived dominance of the Sciences. Naturally, one might use [BC] to describe the

ancient times "Before Computers," knowing that this pair of letters has another specific religious significance for many.

To reinforce said point of intrinsic ambiguity of Capital Letter pairs, one now introduces two characters, of historic note: Plato (Wikipedia Editors, 2014p) and Aristotle (Wikipedia Editors, 2014d). It is these two characters who inspired the strange title of the paper. But, then there is to be a calculated ambiguous twist! Plato is "retired," to be replaced by another: Philo (Wikipedia Editors, 2014o). Naturally, this switch is purely for the benefit of the naming of the chosen paper title, in the first place. However, Philo also occupies an important place in the culture of "Europe."

How, one may ask, does these ancient characters fit into the modern era of computing, of cloud services, of tagging, keywording, ontologizing? For those who have had a traditional humanities upbringing (in Europe, to start with, and extended naturally to the Colonies: Asia, Africa, America, et cetera), it is taken for granted, in the sense that the young (female, male, other) were well and properly educated. Such young people were generally privileged, a counterpoint to the masses of their generation. Similarly, one notes that this also was the case in Eastern Europe, with respect to the Ottoman Empire. Given the location of the conference in Sofia, Bulgaria (Sofia University, 2013), one expects a similar, albeit later development in what may be called the Balkans (a word of Turkish origin, signifying a wooded place). Specifically, on the 24th of May one celebrates "Alphabet Day," more formally known as the "Day of Slavonic Alphabet, Bulgarian Enlightenment and Culture."

It is traditional in many cultures to identify at least one person with whom to associate an important event in the cultural history of a people. For Bulgaria, that person is, perhaps, Paìsiy Hilendàrski (Wikipedia Editors, 2014n).

There is another way to describe the acronym(s) AC/DC. Specifically, it is a universal truth that many concepts are intrinsically ambiguous, that is to say, have more than one meaning (or significance). Given the location of the conference; given the development of Computing in Bulgaria; given the participation mix, it seems reasonable to introduce the original work of an "old Bulgarian," whom the first author met, for the first time, in Varna: Petar Barnev. To get there was a complicated adventure in 1989. The first step was a flight from Dublin to Frankfurt (probably), then onto West Berlin, then a transfer to East Berlin, and finally arrival in Varna. A certain Petar Stanchev picked said author up at the airport, by car (probably not a Lada), and delivered him to the Conference Hotel. He presented the paper, in English, with a live Russian translator (female) of exceeding beauty. It must have been in Varna that he first came across the famous logo shown below in Fig 1.

Specifically, according to Petar Barnev (Markov, 2010), there are 4 activities: Collection, Storage, Processing and Dissemination. These may be paired as [CD] (equivalent to Input/Output) and [SP] (equivalent to Database/WorldWideWeb), the latter being abbreviated WWW. Curiously CD also captures some of the 'banality' of the current WWW. That is to say, there are those who Collect stuff and those who Disseminate that stuff, usually transformed according to the superficiality of basic pretty exposure. The heart and soul of all the real hard work is done in the Storage [Cloud Services, for example, in 2014] and in the Processing [Programmers of Computers]. Naturally, one recognizes the basic input/output in the CD. This corresponds to the human (listen, speak). The other pair, process/store, is clearly related to memory and thus to history. There is a vast amount of input from a variety of sources: images from the eyes, sounds from the ears, feelings from all over the body surface. These inputs may also be stored and processed. (Re)actions may be exhibited, blushing, anger, smiling, indifference, and so on. Such fanciful humanistic language needs to be formalized somewhat.

1.1 Gathering the Information: Barnev's 4 Key Activities

Collection:
It is obvious that organizations such as Google and Facebook are big into collection of data (Krotoski, 2013), (Schmidt, 2014). "The communication technologies we use today are invasive by design, collecting our photos, comments and friends into giant databases that are searchable and, in the absence of outside regulation, fair game for employers, university admission personnel and town gossips. We are what we tweet." (Schmidt, 2014). It is noteworthy to point out that the latter mentions the possibility/likelihood of "a kind of virtual honour killing."p36 Specifically, the context is that of a "deeply conservative society where social shame is held highly," and, for example, in the case of a young woman who has imprudently posted information that ought to have been private, there is a real danger of subsequent "honour killing" whether virtually or actually. There are also efforts in progress to try to limit the scope of private data collected. One such is the "Do Not Track legislation"(Wikipedia Editors, 2014i); another is the "Right to be forgotten (European Union)" (Wikipedia Editors, 2014i).

Storage:
It is taken for granted today that the primary storage of all data is on the Web (and the Internet). In practice, such storage exists physically on Cloud servers. The said servers are kept cool, preferably underground in natural caverns. Typical clouds are those of Amazon (Amazon web services, 2014), Apple (Wikipedia Editors, 2014k), Google (Google, 2014), Yahoo Sherpa (Wikipedia Editors, 2014s) Windows Azure (Wikipedia Editors, 2014r) and so on. It is taken for granted that all Clouds have built in redundancy. It is entirely possible and probable that Cloud servers

intended to cover a specific "Jurisdiction" may be physically located in different "Jurisdictions" to avoid certain kind of "legal difficulties" relating to privacy and so on. This scenario is completely analogous to the issues raised concerning off-shore financial accounts in recent years (Ax, 2014-04-07).

Processing:
Data/information arrives into the system in a variety of forms (text, image, program code, raw binary, etc.) Such data will normally be processed into a standard form, and then depending on the nature of application, further processing will take place. But processing implies programming, and programming implies a programming language (Wikipedia Editors, 2014q). There are a great many programming languages in existence, many of which are rarely used today. One of the latter must surely be the Literate programming language devised by Donald Knuth (Wikipedia Editors, 2014j), (Wikipedia Editors, 2014l). But one needs to ask what exactly is being processed? Today, one might experiment with the Go programming language (Anon, 2014).

Dissemination:
Finally, processed data will be broadcast out into the world. More specifically, the broadcast generally goes out to the World Wide Web (and its users). The current paper "**Re Pub(lic) of Philo(sophy)**" is a typical example. However, one needs to remember that there is the bigger underlying Internet. Collection, storage, processing and dissemination also takes place at this deeper level.

The metallic emblem, shown above in Fig. 1, stands out as a perfect square piece of art in 3 colours: red border (gold lettering), 4 blue panels (2 of which are invaded by golden arrows) denoting Collection, Storage, Processing and Dissemination (in anticlockwise order) and all the rest in gold. It is noteworthy that the emblem is still in common usage with respect to conferences such as (Mac an Airchinnigh, 2014).

Now let us anchor ourselves in the present and ask the fundamental question: will the (near) future be something like the (recent) past? In general, the answer is always "Yes." But let us imagine that June 2014 is radically different? What might still be the same? What might be the catastrophic change? This dramatic outbust allows us to draw the introduction to a close with the mention of Dark and Deep (Internet/Web). How might one discover this for oneself? The question is rhetorical. One just fills out the basic outline.

1.2 The Dark Internet

"A **dark Internet** (Wikipedia Editors, 2014g) or **dark address** refers to any or all unreachable network hosts on the Internet. It is also called dark address

space" (Wikipedia Editors, 2014f).[1] To find more, beyond Wikipedia, one might try a search with Yandex (see below) with search terms: dark internet. Yandex suggests 27 million answers. Here is a rhetorical question! Why choose Yandex and not the usual Google? Does it make a difference?

1.3 The Deep Web

"The **Deep Web** (also called the **Deepnet**, **Invisible Web,** or **Hidden Web**) is World Wide Web content that is not part of the Surface Web, which is indexed by standard search engines." (Wikipedia Editors, 2014h).[2] Once more one is challenged, perhaps even more so, by the existence of the Deep Web. What is it? Why is it there? What can one say about it? Clearly, just like the Dark Internet, there some things which cannot be published in open fora.

2. Seek and ye shall find[3].

What might we be able to do for ourselves without the use of Search Engines? Probably not much! Which Search Engines ought one to use in a controlled experiment to determine access to desired information and how ought one to verify the validity of said information? As a first step we might choose the following, in reverse alphabetical ordering (for a change).

2.1 Yandex

In a "Slavic world," Yandex (Wikipedia Editors, 2014t) is "Tsar?" Let us begin the (re)search by asking Yandex to tell us something about <Petar Barnev[4]>. The correct identification comes in at number 2. But we know this already, as humans. We have known the man. We expected to find the "correct answer," the "correct man."

2.2 Yahoo

Once upon a time, Yahoo used to preserve the first search engine "AltaVista" (Wikipedia Editors, 2014b). Now we have only Yahoo Answers. Let us see how it performs regarding <Petar Barnev>? No results!

---

[1] There is not much one can do with the topic of the Dark Internet. In essence, it is out of bounds.

[2] If it is not indexed, then it is not discoverable? This topic is also out of bounds.

[3] This title is taken from a Christian New Testament, .

[4] We use the angle brackets to indicate that the search is done on the name Petar Barnev, without quotes, etc.

2.3 Google

Naturally, one presumes that Google is the dominant search engine, no matter the language? It is probably the most dominant with respect to the English language? A standard search will immediately (first hit) give notification of death. (Krassimir Markov, 2010)

2.4 Bing

One wonders why Bing is called Bing! Given the American context of MicroSoft one imagines that perhaps the name is taken from Bing Crosby. (Microsoft, 2014).

2.5 Typical Searches on the Web

According to Aleks Krotoski, "BUPA and LSE found that, in the modern web-enabled age, a typical medical consultation follows this trajectory: 1) you discover a growth, 2) you do a Google search, 3) you believe the first result that confirms your expectations." p.162 (Krotoski, 2013). Similarly, John Naughton (the renowned "Networker" reporter for the Guardian and the Observer) recently reported that "The trouble with our big data obsession is that it will help us make gigantic mistakes"). He was referring to the Nature article which reported that "Google Flu trends had gone astray." Specifically, the issue was simple: "Google doesn't know anything about the causes of flu. It just knows about correlations between search terms and outbreaks."

3. Ontologically speaking

Protégé is perhaps the most significant ontology tool in our current era? Naturally, it is augmented significantly by the CIDOC-CRM. Interestingly, it was from the Friedrich-Alexander, University of Erlangen-Nuremberg, Department of Computer Science (Artificial Intelligence), in cooperation with the Department of Biodiversity Informatics… One can read all about it! (Erlangen CRM, 2014). Nevertheless, it is useful to remember that ultimately one wants to make practical use of it.

3.1 Protégé (Ontology)

"Ontologies are the structural frameworks for organizing information and are used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it." (Wikipedia Editors, 2014m).
One would hazard a guess that Protégé (Stanford) is top-class given it origins.

There is just so much information (Collection a la)… In principle (naturally) one ought to focus on the person (philosopher) first and then turn to the concepts. Were one to begin with the concept then there would be myriad people of all cultures who apprehend (and some of whom created) the said concept. Here one choses the 3 ancients Aristotle, Philo, Plato (not by period, but alphabetically). For Aristotle, one will provide basic commentary: https://en.wikipedia.org/wiki/Aristotle [2014-04-12] [B-class] [Locked]. One wonders to what extent each philosopher informed the other. Such wondering entails being specific about ages (and places, and means of communication).

3.2 Aristotle

Let us begin with Aristotle. There is a good sound online account of Aristotle's Metaphysics (Stanford Encyclopedia of Philosophy, 2000). Curiously, this is the same institution that gave rise to Protégé. Now one zooms in on Aristotle's Categories. Essentially one is zooming in to the heart of "ancient ontology." From the Stanford account one identifies that heart: substance (*ousia*). The question now to be resolved is to which part of the CIDOC-CRM shall attach Aristotle's *substance*? Perhaps one might choose 'E63 Beginning of Existence'? Why not? Referring back to the CRM, we are rewarded with 5 core concepts: 'E12 Production', 'E65 Creation', 'E66 Formation', 'E67 Birth' and 'E81 Transformation.' Would not Aristotle be pleased?

3.3 Plato

Naturally, Plato predates Aristotle? The *idea* of introducing Plato is to point, not only to the extensive range of his thinking but to introduce the "concept" of his forms and/or ideas. In particular, one seeks to locate them within the CIDOC-CRM ontology. Given much of the pragmatic use of the CIDOC-CRM, it seems that Plato is captured completely and adequately by the very first concept: 'E1 CRM Entity.' Consequently, everything is included within the Platonic sphere.

3.4 Philo

Philo belongs to a world, very different from Aristotle and Plato. As expressed in the introduction, he is introduced purely for the design of the quirky title. The Philo, in question, is that Philo of Alexandria, a Hellenistic Jewish philosopher (Wikipedia Editors, 2014o). Naturally, he can be squeezed into the ontology. But it would be nice if one could find an idea, a concept, fitting for the paper. Philo was fond of allegory (Wikipedia Editors, 2014a). It would seem that allegory fits nicely into the CIDOC-CRM as the concept: 'E65 Creation,' capturing something of the nature of the person he was.

4. Conclusion

"If we are on the web we are publishing and we run the risk of becoming public figures—it's only a question of how many people are paying attention, and why." (Schmidt, 2014) p. 56.

The paper has focused essentially on what may be considered precise and formal, to a certain extent. Technical language (such as might be required in an ontology) is introduced for that purpose. On the other hand stories have been told, to balance the formality. Every ontology must be explained, interpreted, exhibited for humans. Schmidt's remark cited above is a note of caution. Just because we have mastered the technology, does not mean that our humanity is advanced. One major conclusion is this. The technology needs to be storified. We all need new stories for our computing times.

References

Amazon web services. (2014). http://aws.amazon.com/ec2/ [2014-04-02]

Anon. (2014). The Go Programming Language. golang.org

Ax, J. (2014-04-07). Texas tycoons hid $550 million in profits offshore, U.S. tells jury. http://www.reuters.com/article/2014/04/03/us-sec-wyly-idUSBREA321UG20140403

Erlangen CRM. (2014). CIDOC Conceptual Reference Model. 2014, from http://www.cidoc-crm.org/official_release_cidoc.html [2014-04-11]

Google. (2014). Google Cloud. https://cloud.google.com [2014]

Krassimir Markov. (2010). [ITHEA ISS] IN MEMORIAM: Professor Petar Barnev. from http://www.ithea.org/pipermail/ithea-iss/2010-April/000132.html [2014-04-05]

Krotoski, A. (2013). *Untangling the Web, What the Internet is doing to you*.: Faber and Faber Limited.

Mac an Airchinnigh, M. (2014). Education and Research in the Information Society. http://adis.org/Conference_2014/index-en.html

Markov, K. (2010). In memoriam: Professor Petar Barnev. http://www.ithea.org/pipermail/ithea-iss/2010-April/000132.html

Microsoft. (2014). Bing. Retrieved from https://http://www.bing.com/ [2014-04-05]

Schmidt, E., Cohen, Jared. (2014). *The New Digital Age, Reshaping the Future of People, Nations and Business*.: John Murray.

Sofia University, S. K. O. (2013). Celebrations for 24th May, the Day of Slavonic Alphabet, Bulgarian Enlightenment and Culture. https://http://www.uni-sofia.bg/index.php/eng/news/archive/archive_of_hot_news/celebrations_for_24th_may_the_day_of_slavonic_alphabet_bulgarian_enlightenment_and_culture [2014-04-03]

Stanford Encyclopedia of Philosophy. (2000). Aristotle's Metaphysics. http://plato.stanford.edu/entries/aristotle-metaphysics/ [2014-04-12]

Wikipedia Editors. (2014a). Allegory. https://en.wikipedia.org/wiki/Allegory [2014-04-12] [C-class]

Wikipedia Editors. (2014b). AltaVista. https://en.wikipedia.org/wiki/AltaVista [2014-04-05] [start-class]
Wikipedia Editors. (2014c). Analog computer. http://en.wikipedia.org/wiki/Analog_computer [B-class][2014-03-26]
Wikipedia Editors. (2014d). Aristotle [384 – 322 BCE]. https://en.wikipedia.org/wiki/Aristotle[2014][B-class]
Wikipedia Editors. (2014e). Computer (originally Digital Computer). https://en.wikipedia.org/wiki/Computer [2014][C-class]
Wikipedia Editors. (2014f). Dark Internet. https://en.wikipedia.org/wiki/Dark_Internet [2014][unassessed]
Wikipedia Editors. (2014g). Darknet (file sharing). https://en.wikipedia.org/wiki/Darknet_(file_sharing) [2014-04-05]
Wikipedia Editors. (2014h). Deep Web. https://en.wikipedia.org/wiki/Deep_Web [2014][start-class]
Wikipedia Editors. (2014i). Do Not Track legislation. https://en.wikipedia.org/wiki/Do_Not_Track_legislation [2014-04-02][C-class]
Wikipedia Editors. (2014j). Donald Knuth. https://en.wikipedia.org/wiki/Donald_Knuth[2014][C-class]
Wikipedia Editors. (2014k). iCloud. https://en.wikipedia.org/wiki/ICloud [2014][C-class]
Wikipedia Editors. (2014l). Literate programming. https://en.wikipedia.org/wiki/Literate_programming[2014][C-class]
Wikipedia Editors. (2014m). Ontology (information science). https://en.wikipedia.org/wiki/Ontology_(information_science) [2014-04-11][C-class]
Wikipedia Editors. (2014n). Paìsiy Hilendàrski https://en.wikipedia.org/wiki/Paisius_of_Hilendar [2014-04-03] [start-class]
Wikipedia Editors. (2014o). Philo. https://en.wikipedia.org/wiki/Philo [2014][B-class]
Wikipedia Editors. (2014p). Plato 428/427 or 424/423 BC[a] – 348/347 BC. https://en.wikipedia.org/wiki/Plato
Wikipedia Editors. (2014q). Programming language. https://en.wikipedia.org/wiki/Programming_language [C-class][2014]
Wikipedia Editors. (2014r). Windows Azure. https://en.wikipedia.org/wiki/Windows_Azure [2014-04-02][unassessed]
Wikipedia Editors. (2014s). Yahoo Sherpa. https://en.wikipedia.org/wiki/Yahoo_Sherpa [2014-04-02][unassessed]
Wikipedia Editors. (2014t). Yandex. https://en.wikipedia.org/wiki/Yandex [2014-04-05]

Large Scale Analytics with Hadoop

Vladimir Dimitrov

Faculty of Mathematics and Informatics, University of Sofia,
5 James Bourchier Blvd., 1164 Sofia, Bulgaria
cht@fmi.uni-sofia.bg

Abstract. Big data emergence provokes the great interest on big data analytics tools. Pioneer in this area is Google, who with its solutions influenced the open source project Hadoop. This paper is an overview on Hadoop.

Key words: *Big data, data analytics, Hadoop, HDFS.*

1 Introduction

The term **large scale analytics** is used for advanced analytics techniques applied to big data. This paper discusses what data analytics is and how it works in the context of big data.

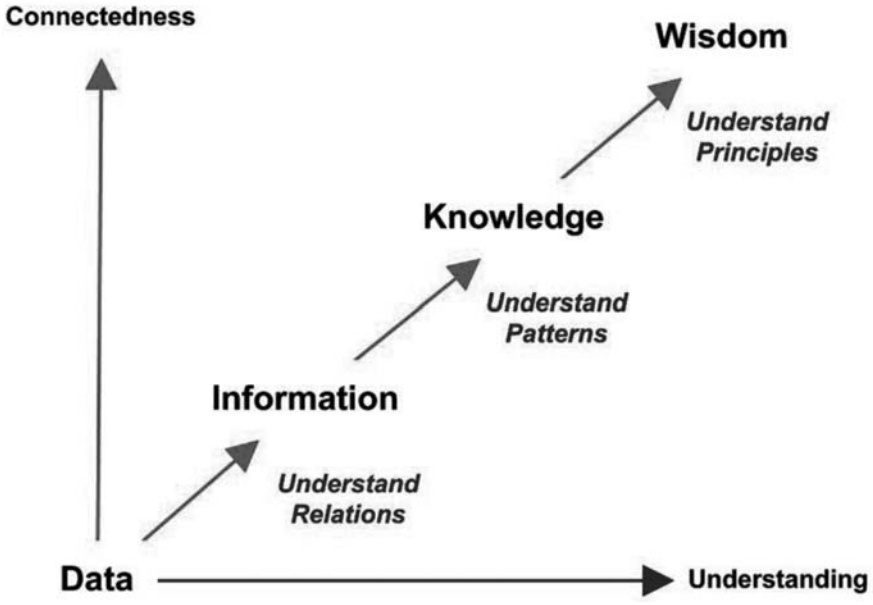

Fig. 1. Data Information Knowledge Wisdom hierarchy.

Information is data that is shaped into a form useful and meaningful for humans. **Data** are streams of raw facts representing events occurring in the

organization or in its environment [6]. The process of data shaping into information is called **data analyses** (**data analytics**). Data, Information, Knowledge, and Wisdom (DKIW) hierarchy defines three levels data processing: Information is retrieved from Data when relations are understood; Knowledge is retrieved from Information when patterns are understood; Wisdom is retrieved from Knowledge when principles are understood as it is shown in Fig. 1. At every level connectedness grows with data understanding.

Russom [8] defines **advanced analytics** as a collection of related techniques and tools used for data analyzes. Advanced analytics includes predictive analytics, data mining, statistical analysis, complex SQL, data visualization, artificial intelligence, natural language processing, and database capabilities. Advanced analytics is also called **discovery analytics** because it discovers information, knowledge and wisdom from the data. **Big data analytics** is advanced analytics techniques applied on big data.

Why big data analytics is so popular? Big data are enormous source of usage samples that can be used for new business models and big data handling tools are available today.

2 What is Hadoop?

Apache Hadoop is the driving force behind today big data industry.

Apache Hadoop is an open-source Java-based framework. It stores data (Hadoop Distributed File System) and executes jobs (MapReduce) on large clusters of commodity servers. Hadoop supports a high-level of fault tolerance. This framework is very simple but effective for a large class of big data applications. It is scalable from a single server to thousands of servers.

Nowadays, research on big data is conducted mainly by the industry. This means that main results are achieved in projects for solving concrete problems. There is no systematic approach to big data. The main considerations in big data analytics can be defined by the following questions are:

1. How big data is stored?
2. How big data is analyzed?

Traditional database systems have limitations on the number of columns, table size, etc. Usually, they require data to be preformatted before to load them in the database. Big data, usually, are not formatted. A big data file could store many terabytes and has billions of fields. A special storage for big data is needed.

Traditional advanced analytics tools are based on relational model of data and N-dimensional cubes. The first approach is to extract data from the big data storage and format them for input to traditional analytic tools. The second approach is to develop big data analytic tools directly on big data storage. It is possible, data extraction to be combined with data processing.

Hadoop combines big data storage and big data analytics. Hadoop Distributed File System is the solution of big data storage and it is discussed here. Hadoop MapReduce is the solution of big data processing.

A solution to above mentioned problems is Hadoop. Usually, when a new technology emerges, there are no standards, but only leading solutions, projects, products, practices etc. Such a leading project in the big data analytics is Hadoop. It combines together data storage and data analytics. Hadoop Distributed File System (HDFS) is the Hadoop's solution of the first problem (Big data storage) and it is discussed here. Hadoop MapReduce is Hadoop's solution of the second problem.

3 Brief History of Hadoop

Doug Cutting [2] created Hadoop. Hadoop is not an acronym. It does not mean anything.

Mike Cafarella and Doug Cutting started Apache Nutch project in 2002. Their idea was to create Web search engine that could index and search one billion Web pages. This engine had to be deployed on half million dollars hardware and with running monthly costs of $30 000 [1]. The main problem in this project was how to store such a big index. Meanwhile in 2003, Google engineers published a paper about Google File System (GFS) [5]. This paper inspired Mike Cafarella and Doug Cutting to create Nutch Distributed Filesystem (NDFS) in 2004.

In 2004, Google engineers Jeffrey Dean and Sanjay Ghemawat published a paper about MapReduce [3]. This paper influenced Nutch project developers and in 2005, MapReduce was available for Nutch.

In 2006, Hadoop was established as a subproject and in 2008 as a top level project in Apache. The developers realized that the project solutions could be used in broader area than Web search.

Nowadays, Hadoop is used by many companies among which are Yahoo! and Facebook. In 2013, Hadoop is the ultimate lieder with 1.42 TB/min (http://sortbenchmark.org/) sorting.

But what in reality is Hadoop? Web search engine, sort utility, file system or something else? This paper tries to give an answer. The next section explains what Hadoop is, for what it is useful and for what is not.

Broadly speaking, Hadoop is an open source implementation of Google File System, Google MapReduce, and Google BigTable etc. Google engineers teach, outside the company, on Google technologies using Hadoop.

4 Hadoop Positioning

Moor's law is still in power: power of electronic components doubles their every 2-3 years. Particularly, disk electronics doubles its transfer rate every 2-3 years, but disk mechanics do not do that in this rate. Standard disk capacity has reached

terabyte level. The time for reading all data from a disk is now more than 30 times longer than 20 years ago. There are some strategies to fight this problem. RAID controllers are an example of successful combination of these strategies. Detailed discussion is available in [4].

One of above mentioned strategies is to use several small disks as a storage instead of one big disk. When the data set is stored on several disks then read/write operation can be performed on all disks in parallel. The longest time in I/O operation is the disk latency, i.e. the time for positioning the package of disk heads on the disk cylinder (track) and then positioning on the track. If the data are stored on consecutive cylinders the disk latency is minimized for all data reading. In this case, the whole data could be read several times faster (linear dependence of number of used disks). There are more benefits in using RAID technology. For example, a file bigger than the capacity of a disk could be stored on several disks. Enormous fault tolerance could be achieved by data replication on several disks etc. See RAID 4 – 6 for more details.

Google File System (GFS) is implemented with some of the RAID controller strategies. The main difference is that RAID controller uses several disks on the same computer system, but GFS uses many disks on different computer systems. Why it is sensible? Because today computer networks could be very fast – to read data from main memory of one computer to another is usually several times faster than to read that data from the local disk. Even more, to read data from SAN system usually is faster than to read that data from the local disk. So, RAID strategies could be applied on several computer systems connected with a fast computer network in the same way as they are applied on several disks connected to a disk controller.

HDFS, at this time, uses two strategies: multiple disks and data replication. The first strategy benefits only when the whole data is read in predictable way (usually in sequential manner). Classical example of such predictable whole data read/write is two/multi-phase sort/merge algorithm. The second strategy (multiple disks) is implemented in very simple way, but it application increases very much fault tolerance even when is used on commodity hardware.

But how big data are processed? Big data are written once and are read many times. Big data analytics reads all or almost the whole big data set. HDFS is suitable for big data analytics. It is optimized to write one very big files and then to read them many times. Usually, big data processing is simple: some data have to be extracted and then evaluated in some manner – there are no long running calculations on every piece of data.

HDFS is not suitable for random reads/writes. Hadoop complements modern DBMS that are based on Object-Relational model of data. Traditional DBMSs are optimized for random I/O operations.

Hadoop is not suitable for Grid computing. Modern Grids are implemented

with the idea of two kinds of nodes (hosts): computing and storage ones. Usually, data is transferred from storage nodes to computing nodes. Grid data processing is long running. For example, a typical bioinformatics task is usually processed on one computing node for several days. Data nodes store huge amounts of data. Data transfers among these two kinds of nodes are not very intensive.

Hadoop does not differentiate hosts in Grid way – one host can run several computing nodes (MapReduce tasks) and several data nodes. Computing nodes, usually, take their inputs only from local data nodes. This is called **data locality organization**.

Hadoop is not stream database system. It does not support unlimited streams. One of the big data sources are sensor data. These are formatted data, but their files are huge. Big data sensor files are not streams. The last ones are endless; they are stored in specialized structures and they are updated all the time. Big data files are written once. Big data sensor data files are finite ones – usually for a given period of time. They are huge but finite and static. Leading DBMS support stream processing option.

Hadoop is open to Clouds. It is designed to be used on commodity hardware (servers, networks). This means that it is easy to be run on cluster of virtual machines in a cloud. Hadoop is easy scalable for large clusters. It is opened to use underlying cloud infrastructure like cloud distributed file system services (abstract distributed file system interface), to achieve higher level of optimization. That is why it is available on the leading clouds, like Yahoo!, Amazon, and MS Azure etc.

5 Hadoop Components

Hadoop is licensed under the Apache License 2.0. It components are:

- Common – utilities that support the other Hadoop components and interfaces to abstract distributed file system.
- MapReduce (YARN) – framework for job scheduling and cluster resource management; programming model and execution engine running on clusters of commodity servers.
- Hadoop Distributed File System (HDFS) – distributed file system running on clusters of commodity computers.
- Avro – serialization system for efficient, cross-language RPC.
- Sqoop – tool for transfer of data between structured data and HDFS.
- ZooKeeper – distributed, highly available coordination service.
- Oozie – service for running and scheduling workflows of Hadoop jobs.
- Pig – data flow language and execution engine for big data.
- Ambari – web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters.
- Hive – distributed data warehouse.

- HBase – distributed, column-oriented database.
- Cassandra – scalable multi-master database with no single points of failure.
- Chukwa – data collection system for managing large distributed systems.

Hadoop core are Common and MapReduce. The current version of MapReduce is MapReduce 2.0 (MRv2) or YARN. All other components are optional. Some of them tend to be implemented in the Hadoop core. HDFS plays special role in Hadoop: it is Hadoop distributed file system, optimized for MapReduce jobs. All these optional components are discussed here.

MapReduce is a framework for parallel processing of large data sets. It originates from functional programming. MapReduce has very simple programming model, but it is possible useful programs to be written in. There are two phases in MapReduce: Map and Reduce. For every phase, the programmer supply a function with predefined interface. In Map phase, the user-defined function is applied on a key-value pair and generates a list of key-value pairs. Then, in Reduce phase, all lists of key-value pairs are sorted in lists of key – values list, where the values list contains all values generated in Map phase for a specific key. In Reduce phase, user-defined function iterates on key – list pair and generates a list of values. This processing can be pure functional (without side effects) and highly parallel in both phases.

6 Conclusion

Hadoop has approved application areas. The question now is what in reality the extent of its application area is. The answer is not very clear because this is a new emerged technology. An open question is the extent of Hadoop application area. Another open question is on the analytics tools for big data. Is there a need to develop specific tools for big data or simply to extract the data from big data storage and load in currently available tools? Many solutions are devoted on the second approach.

Undoubtedly: big data are here, they contain valuable information and their analyses are a big challenge postulating new decisions.

Acknowledgements. This paper is supported by Sofia University "St. Kliment Ohridski" SRF under Contract 05/2014.

7 References

1. Cafarella, M., and D. Cutting. 2004. Building Nutch: Open Source Search. *ACM Queue*, (April): http://queue.acm.org/detail.cfm?id=988408.
2. Cutting D., 2012. Intro to Hadoop and MapReduce. https://www.udacity.com/course/viewer#!/c-ud617/l-306818608/m-312934728.
3. Dean J., and S. Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters. Google (December): http://labs.google.com/papers/mapreduce.html
4. Garcia-Molina, H., J. D. Ullman, and J. Widom. 2009. *Database Systems: The Complete Book (2nd Edition)*. Prentice Hall.
5. Ghemawat, S., H. Gobioff, and S.-T. Leung. 2003. The Google File System. Google (October): http://labs.google.com/papers/gfs.html
6. Laudon, K.C. and J. P. Laudon. 2007. *Management Information Systems - Managing the Digital Firm 10th Edition*. Upper Saddle River, New Jersey: Pearson Education, Inc.
7. Ranieri, B, C. Lucchese and G. De Francisci Morales. 2010. Large-scale Data Analysis on the Cloud. http://melmeric.files.wordpress.com/2010/05/large-scale-data-analysis-on-the-cloud-roma-cmg-2010.pdf (accessed March 3, 2014).
8. Russom, Ph. 2011. *Big Data analytics*. Renton, Washington: TDWI Research.

Distributed Coordination with Apache ZooKeeper

Daniel Simeonov, Vasil Georgiev

Faculty of Mathematics and Informatics
University of Sofia St.Kliment Ohridski"
Corresponding author: dsimeonov@gmail.bg

Abstract: Building coordination algorithms for distributed systems is hard. Nuances like synchronicity and fault-tolerance will make such implementation hard to reason and test. Even if a distribution application implements its coordination that the wheel is invented again and again. With the prevalence of cloud computing it is necessary that applications focus on their business logic and can reuse distributed coordination algorithms. ZooKeeper is a wait-free replicated service which provides API for building complex distributed coordination primitives at the client. ZooKeeper's event-driven mechanism along with the guarantees it offers provides for robust and efficient implementations of distributed constructs like - distributed locks, queues, rendezvous, leader-election, group membership, naming service and other.

Keywords: municipal administrative activities, content management system, information system.

1 Introduction

Building a distributed system is not a simple problem; it is very prone to race conditions, deadlocks, and inconsistency. Making distributed coordination fast and scalable is just as hard as making it reliable. Distributed applications run on different machines and need to see configuration changes and react to them. To make matters worse, machines may be temporarily down or partitioned from the network. Not only do these outages make things hard to configure, but they also make application health no longer a choice between dead or alive; you also have mostly alive or dead and the dreaded half dead. To make matters worse theoretical results such as the FLP proof [2] (consensus is impossible with asynchronous systems and even one failure) and the CAP theorem [3] (strong Consistency, high Availability, and Partition-tolerance: pick two, you can't get all three) mean that some compromises must be made. Besides these theoretical problems there is a set of assumptions architects and designers of distributed systems are likely to make, which prove wrong in the long run - resulting in all sorts of troubles.

According [1], the fallacies are summarized below:

a) the network is reliable;

b) latency is zero;
c) bandwidth is infinite;
d) the network is secure;
e) topology doesn't change;
f) there is one administrator;
g) transport cost is zero;
h) the network is homogeneous.

In this paper, we make an overview of Apache ZooKeeper [4]. With ZooKeeper, these difficult problems are solved once, allowing you to build your application without trying to reinvent the wheel. ZooKeeper is a replicated synchronization service with eventual consistency. It is robust, since the persisted data is distributed between multiple nodes (this set of nodes is called an "ensemble") and one client connects to any of them (i.e., a specific "server"), migrating if one node fails; as long as a strict majority (quorum) of nodes are working, the ensemble of ZooKeeper nodes is alive.

The Zookeeper coordination service does not implement specific higher-level primitives on the server side, but instead it exposes an API that enables application developers to implement their own primitives. This approach enables multiple forms of coordination adapted to the requirements of applications, instead of constraining developers to a fixed set of primitives. This API represents a simple wait-free data objects organized hierarchically as in file systems. Blocking primitives for a coordination service can cause, among other problems, slow or faulty clients to impact negatively the performance of faster clients.

Higher-level coordination primitives can be built on top of the ZooKeeper's API, there include (but not restricted to) – group membership, barriers, distributed locks, leader election, naming service (better variant than DNS), producer-consumer queues, failure detection, rendezvous, etc.

2 Basic Concepts of ZooKeeper

ZooKeeper runs on a cluster of servers called an ensemble – Fig. 1. Besides that a Client API is included for clients to connect to the ensemble. Clients connect to a single ZooKeeper server. Every client maintains a single TCP connection to a single server from the ensemble through which it sends requests and heart beats. If the TCP connection to the server breaks, the client will connect to a different server. The data stored in ZooKeeper is replicated over a set of machines that comprise ensemble. These machines maintain an in-memory image of the data (data is replicated to all of the machines) along with a transaction logs and snapshots in a persistent store. Because the data is kept in-memory, ZooKeeper is able to get very high throughput and low latency numbers. The downside to an

in-memory database is that the size of the database that ZooKeeper can manage is limited by memory.

Clients send read or update requests to the ZooKeeper ensemble. One start-up one of the servers is elected as leader. If the leader server fails a new leader is elected automatically among the other servers.

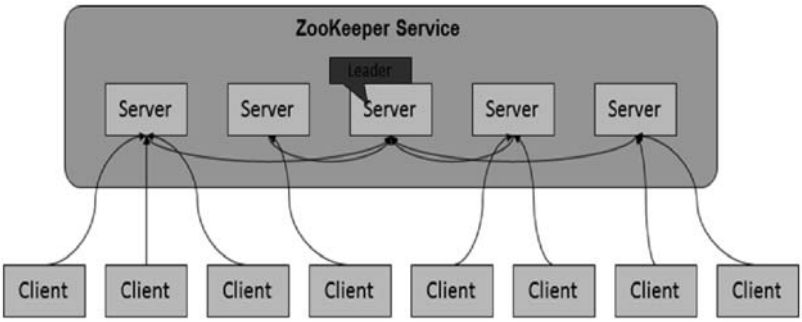

Fig. 1. Process Architecture of a ZooKeeper Cluster.

All update requests are forwarded to the leader. The rest of the ZooKeeper servers, called followers, receive message proposals from the leader and agree upon message delivery. This is done through the use of ZAB, an atomic broadcast protocol. In this way writes can be guaranteed to be persisted in-order, i.e., writes are linear. One the other hand reads are concurrent since they are served by the specific server that the client connects to. However, this is also the reason for the eventual consistency: the "view" of a client may be outdated, since the master updates the corresponding server with a bounded but undefined delay. ZooKeeper guarantees that writes from the same client will be processed in the order they were sent by that client. This guarantee, along with other features discussed below, allow the system to be used to implement locks, queues, and other important primitives for distributed queuing. ZooKeeper exposes lower-level primitives that applications use to implement higher-level primitives. ZooKeeper data model and API resembles a file system with a subset of the operations originally offered by traditional file systems and adds ordering guarantees and conditional writes.

Whenever a change is made, it is not considered successful until it has been written to a quorum (at least half) of the servers in the ensemble.

A ZooKeeper server will disconnect all client sessions any time it has not been able to connect to the quorum for longer than a configurable timeout. The server has no way to tell if the other servers are actually down or if it has just been separated from them due to a network partition, and can therefore no longer guarantee consistency with the rest of the ensemble. As long as more than half of the ensemble is up, the cluster can continue service despite individual server failures. When a failed server is brought back online it is synchronized with

the rest of the ensemble and can resume service. It is best to run ZooKeeper ensemble with an odd number of server; typical ensemble sizes are three, five, or seven. For instance, if you run five servers and three are down, the cluster will be unavailable (so you can have one server down for maintenance and still survive an unexpected failure). If you run six servers, however, the cluster is still unavailable after three failures but the chance of three simultaneous failures is now slightly higher. With more servers, more failures are tolerable, but with lower write throughput.

3 Communication and Data Model

A wait-free implementation of a concurrent data object is one that guarantees that any process can complete any operation in a finite number of steps, regardless of the execution speeds of the other processes [5]. Zookeeper offers wait-free synchronization. This means that are no blocking primitives, such as locks. Also slow or faulty clients do not impact negatively the performance of faster clients. Wait-free property is not sufficient for distributed coordination (compared with blocking primitives like locks for example). But combined with order guarantees for operations and it is sufficient to implement coordination primitives of interest to applications. In particular these guarantees are - FIFO client ordering of operations and linearizability [6] of all writes requests.

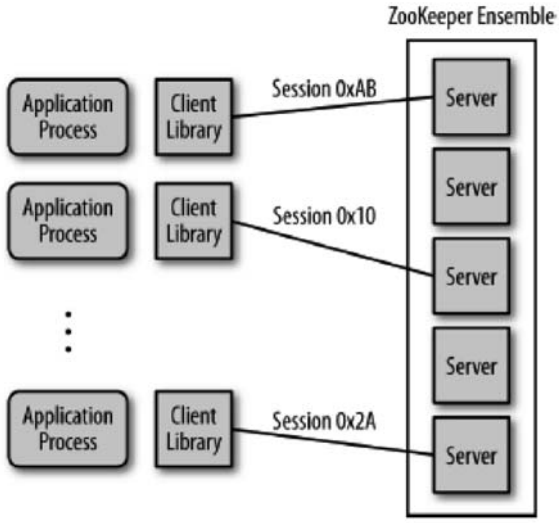

Fig. 2. Client-Server Session in ZooKeeper.

Every client has session on one server – Fig. 2. Clients issue automatic keep-alive (heartbeats) requests with its server. If a client disconnects from a server due

to a timeout, the client tries to automatically connect to another ZooKeeper server while preserving the session. If the disconnection happens because the client has been partitioned away from the ZooKeeper ensemble, it will remain in this state until either it closes the session explicitly, or the partition heals and the client hears from a ZooKeeper server that the session has expired. The ZooKeeper ensemble is the one responsible for declaring a session expired, not the client.

The ***replicated database*** of ZooKeeper comprises a tree of znodes, which resemble roughly file system folders and files. Each znode may contain a byte array, which stores data. Also, each znode may have other znodes under it, practically forming an internal directory system. Each znode also contains version that is incremented every time its data changes. The operations to update or delete a znode can be executed conditionally. Both calls take a version as an input parameter, and the operation succeeds only if the version passed by the client matches the current version on the server.

A znode may be ephemeral: this means that it is destroyed as soon as the client that created it disconnects. This is mainly useful in order to know when a client fails, which may be relevant when the client itself has responsibilities that should be taken by a new client. Taking the example of the lock, as soon as the client having the lock disconnects, the other clients can check whether they are entitled to the lock. A znode can also be set to be sequential. A sequential znode is assigned a unique, monotonically increasing integer. Sequential znodes provide an easy way to create znodes with unique names.

ZooKeeper offers an event system where a watch can be set on a znode. These watches may be set to trigger an event if the znode is specifically changed or removed or new children are created under it. This is clearly useful in combination with the sequential and ephemeral options for znodes. Watches are one time trigger. They have to be reset by the client if interested in future notifications. To receive multiple notifications over time, the client must set a new watch upon receiving each notification. One important guarantee of watches is that they are delivered to a client before any other change is made to the same znode. If a client sets a watch to a znode and there are two consecutive updates to the znode, the client receives the notification after the first update and before it has a chance to observe the second update by, say, reading the znode data.

ZooKeeper data model guarantees:

a) sequential consistency – updates from a client will be applied in the order that they were sent;
b) atomicity – updates either succeed or fail. no partial results;
c) single system image – a single client will see the same view of the service regardless of the server that it connects to;
d) reliability – once an update has been applied, it will persist from that time forward until a client overwrites the update;

e) timeliness – the clients view of the system is guaranteed to be up-to-date within a certain time bound.

4 Higher-level Constructs with ZooKeeper

The ZooKeeper service offers low-level API or primitives with which it is easy to implement more powerful high-level constructs. These constructs are entirely implemented at the client. Some such constructs are not wait-free such as locks (clients need to wait for an event) but they are implemented with low-level wait-free non-blocking primitives. This is possible due to ZooKeeper's ordering guarantees about updates and watches. Watches help to avoid polling or timers but special care should be taken to prevent "herd effect", causing bursts of traffic and limiting scalability. ZooKeeper offers only low-level API primitives which enable the implementation of new primitives without requiring changes to the service core.

Name service and configuration management are two of the primary applications of ZooKeeper. They are easily implemented. Configuration management is easily implemented by storing the configuration in a znode. Processes which would like to read the configuration would read the znode with the watch flag set to true. If the configuration is ever updated, the client would be notified by the watch and will read the latest version of the configuration. Another distributed scenario is group membership - often clients need to know which other processes are currently alive and to get notified upon changes (members die or new members join the group). In this scenario the group is represented by a node. Members of the group create ephemeral nodes under the group node. Clients interesting in the group status would set a watch on the group's znode. Here ephemeral nodes allow these clients to see the state of the session that created the node.

ZooKeeper can be used to implement efficiently distributed locks without the herd effect. These distributed locks include different kind of locks - re-entrant distributed locks, re-entrant read-write locks, distributed semaphores, distributed barriers. Other high-level distributed constructs implementable with ZooKeeper include (but not limited to) - distributed counter, distributed atomic counter, distributed queue (consumer/producer queue), leader election and other.

Distributed leader election is the process where several processes agree between themselves which one is to be leader (only one) and in the case of a failure of the leader the algorithm is executed again to elect a new leader. Leader election is an important and classic problem in fault-tolerant distributed computing with a lot of theoretical research [7, 8, 9]. There are two required properties for leader election algorithms: liveness and safety. Here, liveness would mean "most of the

time, there is a leader", while safety would mean "there are either zero or one leaders".

Here is a sample pseudo-code implementation with ZooKeeper (real implementation in java is much longer to be included here):

```
step1     getData ("/services/myservice/leader", true) //
              here true means that a watch is set on this znode. The watch
              will be triggered upon failure of the current leader and will
              execute this algorithm again; if successful, follow the leader
              described in the data and exit
          create ("/services/myservice/leader", hostname,
              EPHEMERAL) // If successful, lead and exit
          Go to step 1
```

This relatively easy implementation is possible because of the consistency guarantees that ZooKeeper offers.

5 Performance Results

The paper [4] reports performance measurements of ZooKeeper with different read-write ratio and number of servers in the ZooKeeper ensemble. Reads outnumbering writes is typically the case for a coordination service. Performance of read requests is higher than that of write request – reason is that read request are server locally from zookeeper server and are not linearized on the ZooKeeper leader server.

The tests were performed on a cluster of 50 servers. Each server has one Xeon dual-core 2.1GHz processor, 4GB of RAM, gigabit Ethernet, and two SATA hard drives. To simulate larger number of clients one physical machine was used to run several simultaneous clients. The benchmark client uses the asynchronous Java client API, and each client has at least 100 requests outstanding. Each request consists of a read or write of 1K of data. The number of servers that make up the ZooKeeper service varied, but the number of clients stayed the same. The ZooKeeper service was configured in such a way that the elected leader server not to serve read requests.

Scalability of the tested ZooKeeper cluster is presented on Fig. 3. as a number of operations by various number of servers – from 3 to 13. Solid line represents case reading only clients and dashed line represents the case or writing only clients. The linearity of the reads-only case is very good, the write-only case shows very moderate degradation, close to the ideal no-delay for data replication.

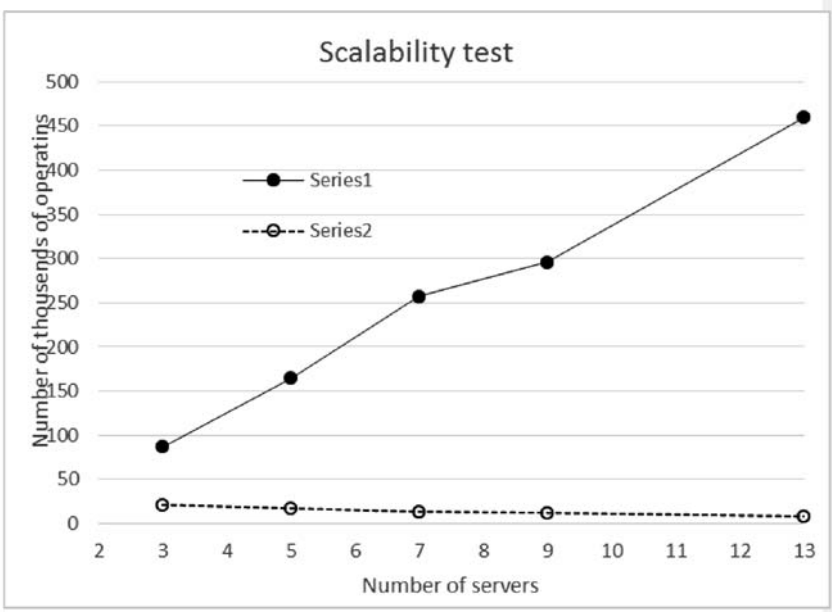

Fig. 2. Number of Operations in a Second by Various Number of Servers.

Reliability benchmark measures the behavior of ZooKeeper as failures are simulated in [4] too. The simulated failures include failure and recovery of a follower, failure and recovery of a different follower, failure of the leader, failure and recovery of two followers and failure of another leader. From this measurement one can observe that if a server from the ZooKeeper's ensemble fail and recover quickly, then ZooKeeper is able to sustain a high throughput despite the failure.

The tested ZooKeeper service was made up of 7 machines with write percentage of 30%, which is a conservative ratio for the expected workloads. The same cluster was in the performance benchmarks.

Acknowledgements This paper is supported by the Project ДДВУ 02-22/20.12.2010 of the National Science Fund.

References

[1] Arnon Rotem-Gal-Oz. Fallacies of Distributed Computing Explained (The more things change the more they stay the same). [Electronic resource], http://www.rgoarchitects.com/Files/fallacies.pdf

[2] Fischer, Michael J; Nancy A. Lynch; Michael S. Paterson (April 1985). "Impossibility of distributed consensus with one faulty process". Journal of the ACM 32 (2). doi:10.1145/3149.214121.

[3] Nancy Lynch and Seth Gilbert, "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services", ACM SIGACT News, Volume 33 Issue 2 (2002), pg. 51-59

[4] P. Hunt, M. Konar, F. P. Junqueira, and B. Reed, "ZooKeeper: Wait-free coordination for Internet-scale systems," in USENIX ATC'10: proceedings of the 2010 USENIX Annual Technical Conference. USENIX Association, 2010.

[5] M. Herlihy. Wait-free synchronization. ACM Transactions on Programming Languages and Systems, 13(1), 1991.

[6] M. Herlihy and J. Wing. Linearizability: A correctness condition for concurrent objects. ACM Transactions on Programming Languages and Systems, 12(3), July 1990.

[7] Hector Garcia-Molina. Elections in a distributed computing system. IEEE Transactions on Computers, C-31(1):47{59, January 1982.

[8] N. Schiper and S. Toueg. A robust and lightweight stable leader election service for dynamic systems. In DSN, 2008.

[9] Lamport, Leslie; Marshall Pease; Robert Shostak (April 1980)."Reaching Agreement in the Presence of Faults". Journal of the ACM 27 (2): 228–234. doi:10.1145/322186.322188.

An Ontology-Based Approach for Integrating of Clinical and Molecular Information for Assistance in Medical Diagnostics

Dimitar Vassilev[1*], Maria Nisheva[2], Nikola Ranchev[3], Velko Ilchev[4]

[1] Agro Bio Institute, Sofia, Bulgaria
[2] Faculty of Mathematics and Informatics, Sofia University "St.Kliment Ohridski"
[3] Medical Information Systems Plovdiv
[4] Technical University Plovdiv
*corresponding author: jim6329@gmail.com

Abstract: The paper presents a work in progress aimed at the design and implementation of an IT platform for integration of clinical and bio-medical (genetic, biochemical, metabolic, and molecular) information in order to provide syntactic and semantic interoperability between various software tools for data and knowledge processing. This platform will be used as a powerful and flexible environment for medical practice and research, in particular for translating genomic and other research information for the purposes of prenatal diagnostics and prevention of rare diseases. The obtained results will also be usable for risk assessment in medical insurance and social health care.

Keywords: information system, knowledge based system, ontology, decision support, data integration, prenatal diagnostics

1 Introduction

The recent rapid progress in technologies and methods for data generation and storage led to pitting of information practically in all domains of research, business and other human activities. In medicine these processes are notably tangible mostly due to the heterogeneous profile of the generated data concerning the health status, diagnostics, medical treatment and testing of the patients. Sources of such information could be medical check records, medical tests, clinical biochemistry profiles, 3D images from computer tomography, biometric data, DNA tests from molecular diagnostics, etc.

All these amounts of collected and stored data significantly exceed the possibilities of using them in an effective and proper manner without a specially designed infrastructure for handling and analysis. As a result the data collected in large data bases looks like a rarely accessed big archive vault. This situation could be described as "rich in data and poor in information". It can be vanquished by the development of suitable ICT infrastructure providing solutions for information retrieval, data mining and knowledge discovery.

Latterly an obvious improvement in funding, managing, and development of information systems in medical practice, medical research and health care gains more power. Medical checks and treatment more and more rely on the information produced by the contemporary technical equipment for molecular analyses such as high-throughput sequencers, mass-spectrometers, chromatography, microarrays, tomography, echo-graphs and other lab equipment. The contemporary medical diagnostics is based not only on usual clinical data of the patient but also on various biochemical, cellular, DNA, proteomic, metabolite and other analyses. This data is generated from clinical patient dossiers, pre-clinical medical records, hospital laboratory test cards, archives, extra diagnostics records, etc. All this requires the development of innovative IT infrastructure for bio-medical purposes oriented towards integration of clinical and other medical research data, information retrieval, data mining, visualization, data security, privacy and reliability, search knowledge retrieval [1]. The final aim of such development is the improvement of diagnostic tools based both on the clinical records of the patients and the translation of research information in medical practice.

A large and famous project in this domain is the I2B2 [2], which is funded by the NIH with Roadmap National Centres for Biomedical Computing [3, 4]. This project provides to clinical and research medical staff a set of software solutions for effective integration of clinical and bio-medical information of various origin. In the EU there are several projects for integration of clinical and bio-medical data for the purposes of the translational medicine. The ONCO-i2b2 platform is a suite of bioinformatics tools designed to integrate clinical and research data and to support translational research in oncology. It is implemented by the University of Pavia and the IRCCS Fondazione Maugeri hospital (FSM), and grounded on the software developed by the Informatics for Integrating Biology and the Bedside (I2B2) research centre. I2b2 has delivered an open source suite based on a data warehouse, which is efficiently interrogated to find sets of interesting patients through a query tool interface [1]. Similar are the purposes of the project EuroGenTest [5]. EuroGenTest is an EU-funded Network of Excellence (NoE) with 5 units looking at all aspects of genetic testing - Quality Management, Information Databases, Public Health, New Technologies and Education. The EuroGenTest Clinical Utility Gene Cards (CUGC) are disease specific guidelines dealing with clinical utility of genetic testing, the ability of genetic test results to reveal information essential for the clinical setting. Information concerning the clinical testing is divided in: differential diagnostics, predictive testing, risk assessment in relatives, prenatal testing. Also in the EU the ELIXIR (European Life Sciences Infrastructure for Biological Information) project [6] is in progress. ELIXIR is a platform for integration of information and research infrastructures, which will play important role in translation of bio-medical (genomic) data in medical diagnostics.

The development of projects for integration of clinical and research biomedical information aims both the improvement of medical diagnostics and also prevention of some social diseases and betterment of health care quality. These targets focus not only the particular patient but consider in a large scale the conception of development of electronic health record standards both in every EU member country and in the whole EU.

A crucial role in these efforts plays the regional development and implementation of suitable ICT platforms for integration of clinical and biomedical information. Such platforms has also a real capacity for providing data for population studies for the specific diseases in different geographic and population regions, but also provide information for typical local diets, the ecological manner of life, the general practitioner practices, etc.

This paper presents the idea and the methodology of a research project directed to the development of original architecture, algorithmic and implementation solutions for an IT platform for integration of clinical and molecular information in prenatal diagnostics. This IT platform incorporates various software tools for data and knowledge processing.

2 Objectives

The main goal of the project is development of methods and information technologies for creation of a platform for incorporation of clinical, diagnostic, research and other medical information based on the integration of different types of data and provision of adequate web-based access of different categories of users. The platform is intended for implementation in rare diseases and prenatal diagnostics performed in genetics laboratories and centres for molecular medicine.

In details this major purpose comprises the synchronization and amalgamation of the existing information by procedures of mining, alignment, merging, normalization and transformation of the data. The development of a suitable meta-format for data transfer is in progress. The development of subject dependent methods for knowledge discovery from non-structured information sources is among the tasks of the project. A subject ontology and proper methods for classification of symptoms and characteristics of the targeted diseases will be created on the base of semantic integration of clinical tests and molecular diagnostics. Visualization and access to the chosen subject domain is provided in the context of various web-based and mobile technologies, following the necessary requirements for authenticity, privacy and anonymization. The potential for scalability of the platform is intended to provide easy integration of the processed information for the purpose of the e-health patient dossier standards. In this way the intended platform will facilitate not only the clinical and medicine benefits from the developing platform, but will also the medical insurance practice.

The provided work schedule emphasizes the particular importance of the

development of innovative approaches for data integration from different sources and integration of data and knowledge from domains as diagnostics, risk assessment, and prevention of rare diseases as well as from the health insurance sources. The development of an ontology, federating knowledge from the above cited subject domains and the elaboration of methods for medical advises and diagnostics suggestions will contribute the data integration strategy of the project. The defined ontology could be implemented through an expert module which also achieves the applied methodology for decision inference. The included units for statistical analyses and population studies on the integrated clinical and molecular data for the purposes of innovative medical research and diagnostics will contribute also to the translation of the related research results in the medical practice and health care. The architecture of the platform will be open with up-to-date functional capacities for integration of data from new bio-medical sources. This architecture complies not only with the functional options but also with the requirements for secure access, handling and protection to the integrated data. A web-based module for administrated dissemination and distribution of the information is also considered in the system architecture - this module will serve to the purposes of the social health care, health insurance and disease prevention and diagnostics.

3 Methodology

3.1 Information Platform

In the course of design of the architecture and functional facilities of the information platform a study of some similar projects has been carried on. They focused on the specific formats for presentation of clinical and molecular data analysis. The development of the means for data integration by using suitable meta formats is also a sort of research objective of the discussed work. An import/export data format suitable for the transfer from and to existing information systems (LIMS, old clinical and lab data bases) is under development as well.

The current hypothesis for the architecture of the information system is appropriately developed in terms of SOA (Service Oriented Architecture) as it is shown on Fig. 1. The SOAP (Simple Object Access Protocol) cover of the shell will contain the service name and the appropriate data from various test sources and analyses.

Two options for database management system (DBMS) have been discussed at the moment: Oracle Database 11g Standard Edition and PostgreSQL 9.x. Oracle database is the undisputed leader in DBMS. PostgreSQL is recognized as the best open source DBMS with very good functionality, scalability and reliability.

The basic development environment will be Eclipse - it will be used for the development of applications in Java, C++, HTML5/CSS/JavaScript, UML.

The Application server will be Java based and will be developed in Jboss or GlassFish, if we choose Oracle DBMS.

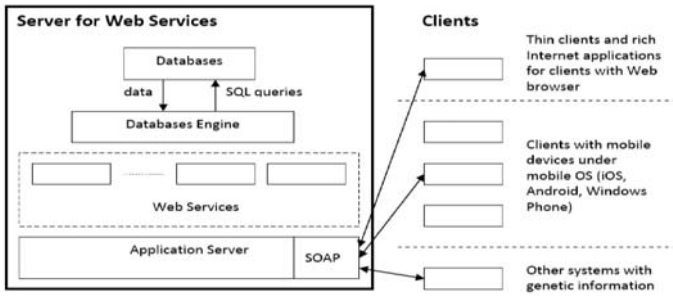

Fig. 1: Information system architecture diagram

3.2 Knowledge Based System

The knowledge based system will have a multi-layer structure. The basic layer will have the form of subject (domain) ontology, while the next layer will consist of a production rule system provided with a proper formalism for uncertain knowledge representation. A third layer with meta-knowledge will be framed if appropriate.

The subject ontology describes the basic domain concepts with their corresponding properties and related logical axioms. The production rules describe the expert knowledge, which will be used in the generation of diagnoses and recommendations. In fact they define the most essential relationships between the declarative knowledge units included in the ontology and the features and values of different kinds of data (clinical data and results of bio-chemical, DNA and other tests). Certainty factors have been used in order to represent the heuristic nature of a part of the domain knowledge. Their range and specific values have been set with the assistance and manual curation of a group of experts in genetics and molecular medicine.

The domain ontology has been under development with the use of the popular open source knowledge editor Protégé [7, 8]. The knowledge model is based on the logical (OWL DL) one supported by Protégé. Most classes of the subject ontology have been constructed as defined OWL classes, by means of necessary and sufficient conditions defined in terms of proper restrictions on certain properties.

In the process of knowledge elicitation, classical textological methods (most of all, analysis of textbooks and other specialized literature) and communicative methods (interviews and brainstorming sessions with domain experts have been used for the purposes of building the particular layers of the knowledge base (Fig. 2).

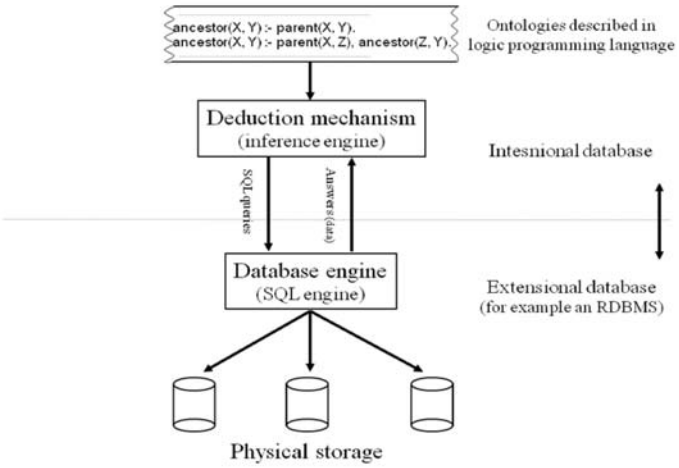

Fig.2: Structure of a WEB-service for processing ontologies

Some of the WEB-services will be able to handle ontologies. As a parallel project, we plan to develop our own description language, to describe specific cases in our subject area. We also plan to develop an interpreter for this language as well as an inference engine which will help us to extract additional knowledge from the facts, stored in a relational database.

Some freely accessible domain ontologies and thesauri such as UMLS [9], FMA [10], WordNet [11] as well as methods and tools for semantic mapping and merging of domain ontologies [12, 13] have been used for knowledge acquisition purposes.

4 Conclusions

The integration of data and knowledge from the subject domain ”prenatal diagnostics, risk assessment and prevention from rare diseases“ will contribute to a long term plan for providing opportunities of new quality in the analysis of the available information and generation of more complex inferences.

The creation of an ontology integrating knowledge in the defined subject domain as well as in the domain “health insurance” will upgrade and unite the multidisciplinary pattern of the collected data and will contribute the elucidation of the logical relations of the data complexity as well as the development of methods for making inferences on the background of the available knowledge and data.

The elicitation of information of new quality will be a cornerstone for many types of analyses, assessments and up-dating of the tendency of the related domains development.

The discussed platform will facilitate the work of experts in medical genetics and molecular medicine and will improve the efficiency of processing and analysis of the available data by different research groups.

The development of a dissemination module for the purposes of the health insurance, social medicine, and for providing information access to the particular patient to medical diagnostics will assist the rapid implementation of e-health services as well as the translation of results and achievements of the molecular research in medical practice giving an opportunity for personalization of the health care and treatment of patients with both frequent social important and rare inherited diseases.

Acknowledgement. The presented work has been partially funded by the Sofia University SRF within the "New approaches and information technologies for building special-purpose knowledge based systems" Project, Contract No. 044/2014.

References

1. Segagni D., Tibollo V., Dagliati A., Zambelli A., Priori S.G., Bellazzi R.: An ICT infrastructure to integrate clinical and molecular data in oncology research. *BMC Bioinformatics **13** (Suppl 4):S5* (2012).
2. Informatics for Integrating Biology and the Bedside, http://www.i2b2.org (accessed on September 1, 2014).
3. Murphy S.N., Mendis M., Hackett K., Kuttan R., Pan W., Phillips L.C., Gainer V., Berkowicz D., Glaser J.P., Kohane I., Chueh H.C.: Architecture of the open-sourceclinical research chart from Informatics for Integrating Biology and the Bedside. *AMIA Annual Symp Proc 2007, 548-552* (2007).
4. Mate S, Bürkle T, Köpcke F, Breil B, Wullich B, Dugas M, Prokosch HU, Ganslandt T: Populating the i2b2 database with heterogeneous EMR data: a semantic network approach. *Stud Health Technol Inform,169:502-506* (2011).
5. EuroGenTest, http://www.eurogentest.org/ (accessed on September 1, 2014).
6. European Life Sciences Infrastructure for Biological Information project (ELIXIR), http://www.elixir-europe.org (accessed on September 1, 2014).
7. Protégé homepage, http://protege.stanford.edu/ (accessed on September 1, 2014).
8. Rubin D., Shah N., Noy N.: Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics, 9(1):75-90* (2008).
9. Bodenreider O.: The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Res., 32, 267* (2004).
10. Rosse C., Mejino J.V.L.: A reference ontology for biomedical informatics: the Foundational Model of Anatomy", *Journal of Biomedical Informatics 36:478-500* (2003).
11. WordNet project homepage, http://wordnet.princeton.edu/ (accessed on September 1, 2014).
12. Petrov P., Krachunov, van Ophuizen E.A.A., Vassilev D.: An algorithmic approach for inferring cross-ontology links wile mapping anatomical ontologies. *Serdica Journal of Computing 6(3):309-332* (2012).
13. Petrov P., Nachev N., Vassilev D., Krachunov M., Nisheva M., Kulev O.: ANATOM: An intelligent software program for mapping and merging of anatomical ontologies. *Proc. of ISGT 2012 Conference*, Sofia, St. Kliment Ohridski University Press, pp. 173-187 (2012).

METHODS AND TECHNOLOGIES FOR EMAIL PROTECTION

Falak Hasan

Faculty of Mathematics and Informatics,
University of Sofia "St. Kliment Ohridski", Bulgaria

Abstract. This paper describes most important email protections technologies, security tools like email protocols, firewalls, Intrusion Prevention System, Intrusion Detection System, other important subtopics like hacking and avoiding hacking, spams and anti-spams all are explained and then there is analyzing and discussion of some researchers' results done in this field, finally an overall description of conclusion and future directions is submitted to demonstrate, searching and figure out robust dependent tools to check security system performance depending on some standards.

Keywords: Client-side attacks, defense in depth, Intrusion Prevention System, filters, spams, anti-spams, hacking, Email services and Email protocols.

1 Introduction

E-mail is a tool to transfer information from sender to exchange information around the world through the internet; it has become an important aspect of the new life and it is an important thing to everyone that cannot complete a work without. Not in the same ratio as researchers and scientists are using in the habit but using email is increased rapidly by everyone respectively.

Robert Shimonski added in [1] that the history of messaging clients such as email were limited in functionality and scope, were generally tied to the operating systems in which they were developed with and did little to supply active content, or any content at all for that matter. The vector for attack was minimal and the ability to penetrate a system or exploit it, limited as well.

To protect clients, firewalls and proxies can be used if contact from local or internet resources on the internet is controlled with these types of defenses tools. It is necessary to know that some technologies must be combined with other powerful technologies like Intrusion Prevention System in order to provide required protection and this method is better than if a firewall works separately. There is another tool called Intrusion Detection System can be defined as a software or hardware component that automates the intrusion detection process. It is designed to monitor the events occurring in a computer system and network and responds to events with signs of possible incidents of violations of security policies [11].

Generally Intrusion prevention system is defined as a device or program used to detect signs of intrusions into networks or systems and take suitable action, as a result alarms can be generated and/or dynamically delaying attacks. Even more robust filtering solutions are available, but typically these only protect a limited set of client technologies. Firewalls are used to filter and protect from the most common attacks [1]. A user may receive an email from either a known or unknown source in these cases and open or preview it normally runs the code to do as the attacker wanted.

The main reason client-side attacks are effective there is a chance of happening lack of defenses on a client. As any security or IT professional can notes, the number of systems that actively use protection such as antivirus is quite low meaning that many systems are left unprotected or without enough protection. The general misunderstanding is that when users became victims of client-side attacks, they supposed they were protected because they installed such software. This software not only needs to be managed and monitored correctly, it also needs to be updated and run often. Sometimes this software itself is compromised given one a false sense of security [1].

Additionally of IDS and IPS tools other technologies are used to protect email from all circumference conditions through the internet like anti-spams, filters, hacking prevention techniques and others are described and discussed in the following sections.

2 Security E-mail Protocols

Security E-mail Protocol in general can be defined as a sequence of operations that confirm protection of data. Used with a communications protocol, it offers secure transmission of data between two parties. It refers to a suite of elements that work consecutively. Nowadays, e-mail has represented as one of the most extensively used communications medium. Since it is low cost characteristics and rapid transfer of messages, e-mail is increasingly used in place of natural mail. However, the e-mail service discoveries users to several risks associated to the weakness of security during the message exchange. Moreover, regular mail provides services which are generally not offered by e-mail, that are of essential importance for “authorized” actions [4].

Certified e-mail tries to provide users with additional guarantees on the content and the delivery of the messages, making e-mail equivalent and in some cases more convenient than the ordinary paper-based mail service. These protocols provide the following property: user Bob receives an email message from user Alice if and only if the latter receives a receipt for this communication, i.e., a proof that the message has been delivered to the recipient. The receipt is such that the recipient cannot deny having received the message. This feature is called

non repudiation. In order to implement non repudiation many of the protocols use a Trusted Third Party (TTP) whose job is to ensure that any disputes which may arise between the communicating parties can be settled in a fair, unbiased manner. TTPs can be of various kinds including inline, offline etc. [2].

Post office protocol 3 is very simple way to access email by email server. Using email clients like mobile Thunderbird, devices, or outlook etc. It downloads also a copy of email existing on user email server with a selection to keep a copy of emails on email server or remove all emails from server. Accessing the downloaded copy of email without an internet connection is the best POP3 feature. The downloaded emails are like SMS's in the inbox, and it can be read anytime without using network. POP3 have a weakness also that it downloads all emails containing spam and viruses.

Internet Message Access Protocol is the most standard and secures protocol to access email from local email server. It is unlike POP3 when a user makes a query for accessing an email, then it downloads a copy of that specific email. It always recalls a copy of emails on email server. This protocol needs lowest data transferal to access the email, so with equal efficiency any slowest modem internet connection will work. Folder creation and deletion functions performed by IMAP also POP3 can perform. But the best feature of IMAP is that it uses encryption while sending and receiving emails, this makes the transmission more secure. Also Gmail uses IMAP protocol for accessing email from email server.

Simple Mail Transport is an application protocol defined in the early 1980s to support email services [10]. It is the best broadly protocol for e-mail delivery. It is very weak in security features for privacy, authentication of sending party, integrity of e-mail message, consistency of e-mail envelop and nonrepudiation envelope. To make e-mail communication secure and private, e-mail servers combine one or more security features using add-on security protocols. This protocol uses encryption which makes email sending and receiving more secure.

M. Tariq Banday added in [3] that previously several technological and policy changes were made to SMTP servers to make e-mail system secure without creating incompatibility between older and newer systems. These include SMTP session refusal to unauthorized servers through IP address verification, refusal of e-mail relaying, restriction on use of certain SMTP commands like EXPN, verification of e-mail envelope and headers, limiting the size of e-mail message and filtering. These security features were updated, upgraded and some of them have been standardized. These security features fall under two broader categories namely technological and legal solutions. Technological solutions include solutions that suggest process or protocol change or use of one or more add-on security protocol or use of some machine learning or non-machine learning filtering technique. In some parts of the globe specific legislative measures are in vogue to deal with legal issues arising from security lacunas of email systems.

3 Viruses versus Anti-Viruses

In the following there are some types of viruses [1]:

- Worms: are programs that can be run independently, they will ingest the resources of its host in order to maintain itself, and can spread a complete working version of itself on to other machines.
- Trojan Horses: the name comes from that critical event in the novel The Iliad, when the Trojans, during the battle of Troy, allowed a gift of a tall wooden horse into the city gates are code disguised as benign programs that then behave in an unexpected, usually malicious, manner. The limitation of Trojans is that the user needs to be convinced to accept/run them.
- Hoaxes: as odd as it sounds, the anti-virus has also taken it upon itself to track the various hoaxes and chain letters that circulate the Internet. While not exactly malicious, hoaxes tend to mislead people; just as Trojan horses misrepresent themselves.

Anti-Viruses are full solutions to almost every existing virus problem, and sometimes solutions to non-existing problems as well. The most popular solution is to regularly scan the system looking for known signatures.

As shown in [12], mailing is one of the interfaces to bring viruses effects.

One of the most important factors in the successful protection of the network against viruses is how fast getting new virus engine signature files – those files released by antivirus labs that help to identify a virus when there is a virus outbreak. Email allows viruses to be spread at lightning speed in a matter of hours; and a single email virus is enough to infect whole network. Clearly then, a critical factor is how fast the signature files of your antivirus solution are updated when a new virus emerges. In every virus attack there is a time differential between the outbreak of the new virus and the release of signatures to defeat and eliminate it. The faster a signature file is created, the less likely the chance of an infection.

4 Hacking Versus Preventing Hacking

In order attackers to get into any network, and once they're in, they truly have the k**e**ys to the kingdom. Not only can they gain higher level access to the network, especially if they launch an elevation of privilege attack, they now can see the entire targeted user's email content, and at the same time can impersonate that user by taking over their identity. And email is far too vulnerable. Not only are passwords commonly weak, but users are easy prey for social engineering, and controlling a user's address book is a bot's delight [5].

4.1 Phishing (or Pharming)

Sending an email to a user and falsely demanding to be an established authentic individual or enterprise in a challenge to force the user into providing private information that will be used for recognizing theft. Such emails usually direct the victim to visit a website where they are fooled into providing or updating personal information, such as passwords, credit card, social security, and bank account numbers, that the legitimate organization already has. However, the website is bogus and set up simply to thief the user's information.

Ignoring the link in the email is the best way to avoid being a victim of a phishing attacks. Altogether it can be bypassed by going to the website in question directly to verify it. Don't submitting any private information online, especially to entrusts sources. Most enterprises will not ever ask to modify personal data like social security number, bank account numbers, credit card numbers or any other sensitive data via email. There is possible of open to possible fraud by supplying any of this information [1].

4.2 Sniffing

Sniffing is defined as an application that can capture network packets. The Sniffers are known as network protocol analyzers and they are really network troubleshooting tools, they used by hackers for hacking network. The data within the network packet can be read using a sniffer, if the network packets are not encrypted, Sniffing is a process used by attackers to chunk the network traffic by using a sniffer. The contents of packets can be analyzed when the packet is chunked by a sniffer, Sniffers are used by hackers to capture sensitive network information: passwords, account information etc.

As a tool of defense against hacking like hoaxes attacks as example is education and common intelligence among the user base not to mention a clear way to report doubtful content as a substitute of blasting it out to everyone under the sun. To simply forward on junk mail there have been many attempts at end users, or to create a DoS attack by resending emails over and over again.

Additional techniques to prevent hacking, after installing strongest filters and anti-viruses, these points must be taken to keep client's privacy information:

1. Improving client's aware and acknowledges not filling forms and styles with real information.
2. Not to remain his email active for long time
3. Using emails with not real information if he has doubts when achieving some works.
4. Building a strong security system with different levels and privileges in the institutes.

The following points must be put in attention by the administrators [9]:

1. Managing switches in as secure a manner

2. The native VLAN ID should not be used for trunking. Using a dedicated VLAN ID always for all trunk ports.
3. Setting all user ports to non trunking
4. Doing configure port-security feature in the switch for more protection.
5. Avoiding using VLAN
6. Deploying port-security anywhere possible for user ports
7. Enabling BPDU Guard for STP attack mitigation
8. Using private VLAN where appropriate to further divide L2 networks
9. If VTP is used, then using MD5 authentication is better.
10. Disabling unused ports.

5 Spams versus Anti-Spams

Email spam known as junk email, it is a subset of electronic spam involving nearly exactly same messages sent to many recipients by email. When Clicking on the links in spam email this may send users to sites that are hosting malware or phishing web sites. Spam email includes malware as scripts or other executable file attachments.

A spammer needs three elements to execute a spam operation: a list of victim email addresses, content to be sent, and a bot- net to send it. Each of these three elements are critical for the success of the spam operation: a good email list should be composed of valid email addresses, a good email content should be both convincing to the reader and evades anti- spam filters, and a good botnet should efficiently sent spam.. Given how critical these three elements are, figures specialized on each of these elements have emerged in the spam ecosystem. Email harvesters craw the web and compile email lists, bot-masters infect victim computers and maintain efficient botnets for spam dissemination, and spammers rent botnets and buy email lists to run spam campaigns. Previous research suggested that email harvesters and bot-masters sell their services to spammers in a prosperous underground economy. No rigorous research has been performed, however, on understanding the relations between these three actors. [6].

On the other side, spammers need an effective infrastructure to sell the illicit goods that they advertise. This infrastructure includes the websites that sell the goods, the shipping facilities, and the payment processors Thus, there can be three main parties involved in the spam ecosystem: the email harvester, the botmaster, and the spammer. Studying the relationship among these different parties involved in the spam ecosystem deepens the understanding of the spam underground economy and can pave the way for new spam mitigation techniques. In this way, it first helps to estimate the magnitude of the spam problem and can reveal new trends. Second, it allows to identify bottlenecks and critical points in

the spamming pipeline; these critical points can be used to develop mitigation techniques to such threats. For these reasons, previous work analyzed individual aspects of the parties involved in the process. In particular, researchers studied the harvesting process of email addresses on the web, the structure and operation of spamming botnets, or the email templates used by spammers. Other work focused on studying the financial conversion of spam or the workflow that goes from when an illicit good is purchased to when it gets delivered. These recent advances in the understanding of individual parties now open the question on their relationship.

To avoid spams a correct Anti-spam have to be Chosen depending on Server-based or client-based, the one that depends on client level needs more time because it requires to deploy it to all workstations on the network, needs updating the anti-spam rules on each of them also means the email infrastructure is being overloaded by spams, as the server message stores are flooding with useless emails it waiting for deletion.

To block spam effectively, we need to have a server-based anti-spam product that offers these advantages:

1. Installation at the gateway eliminates the deployment and administration hassle involved with desktop based products.
2. Far cheaper to license.
3. Prevents spam from even entering your email infrastructure, meaning that your email stores are not full of spam messages.
4. Server-based anti-spam software has more information, and can do more to detect spam effectively.

A mathematical approach based on known spam and valid email used in Bayesian filter. This gives it a tremendous advantage over outdated spam technology that just checks for keywords or relies on downloading signatures of known spam. It provides the following advantages:

1. It is not only keywords or known spam signatures ,also it looks at the whole spam message
2. Not only Learns from the outbound mail also reduces false positives greatly
3. By learning about new spam and new valid mail it adapts itself over time.
4. To the company dataset is unique, making it impossible to bypass Multilingual and international.

6 Firewalls, IPS and IDS

Firewalls are used to filter and protect from the most common attacks. IPS represents a second level of protection to filter and protect from attacks by using heuristic which is a method where signatures are downloaded to an engine in order

to find anomalies in network traffic. When used together they provide "defense in depth." Email represented the exposed user/client main goal information to attacks, [1] where a user receives a message with a malicious payload via a script attached or embedded into it is an example of user's email attack. Intrusion Prevention System [1] is a critical part of an organization's overall network and systems protection strategy and a critical part of a defense-in-depth architecture.

IDS system Security tools such as IDS/IPS are often the only defense in protecting sensitive data and assets besides for being educated when it comes to mitigating attack.

Both firewall and IPS are necessary to provide a robust security system of corporation because each has different capabilities. The differences between them are; the firewall is designed to block all network traffic but that which is obviously allowed while an IPS is designed to permit everything excluding that which is obviously rejected, a firewall is designed to allow or prevent network packets based on their source, destination, and port number regardless of the contents of the message while an IPS is designed to allow or block network packets based on the packet's contents.

Also Robert Shimonski in [1] mentioned that IDS/IPS systems are network security appliances that monitor network and/or system activities for malicious activity. The main functions of intrusion prevention systems are to identify malicious activity, log information about this activity, attempt to block/stop it, and report it.

7 Email Security Clouds

Moving organizations' emails to the cloud, or just wishing to keep their email security at support's length where a third-party authority can maintain and update the solution, the cloud-based email security service is growing in popularity may become an organization's aim.

The buffer between the mail server and the wider Internet is Cloud-based email security. Before all inbound and outbound email being delivered to the mail server, it is received at the security service, whether that server is also in the similar cloud, a different cloud, or even back on the evidences. Doing this ensures that the content is virus-free and confirms with content policy before it is released for sending or for downloading to a client PC [4].

This buffer approach delivers a range of end user and IT department benefits:

- Virus scanning for Inbound and outbound
- Spam filtering before the point of reception
- Administration is easy
- Both maintenance and updates are centralized
- Point of failover

Other positive points also can be found:

- There are operational and acquirement benefits to the cloud approach, such as requiring no physical hardware to be purchased or maintained on site.
- As the business requires, capacity can be scaled up or down and also no software licenses to buy,
- Paid-for licenses need no longer be exhausted or left unused due to a contraction in the number of users.

In a period of constant growth Cloud-based security is suitable for the companies and in this state, the work may select to have the email server hosted on principle for security and agreement purposes however to avoid having to deal with licensing of an email security solution and maintenance, it chooses a cloud-based service that takes care of its email security needs. Java Message Service (JMS) is an example of middleware of the Java Platform for sending messages between two or more clients [10].

As shown in Figure 1 [10], email service in (a) both sender and the receiver communicate asynchronously using inboxes and outboxes. Mail demons run at each site. In (b) an event service supports coordination in a distributed system environment. The service is based on the publish/subscribe paradigm; an event producer publishes events and an event consumer subscribes to events. The server maintains queues for each event and delivers notifications to clients when an event occurs.

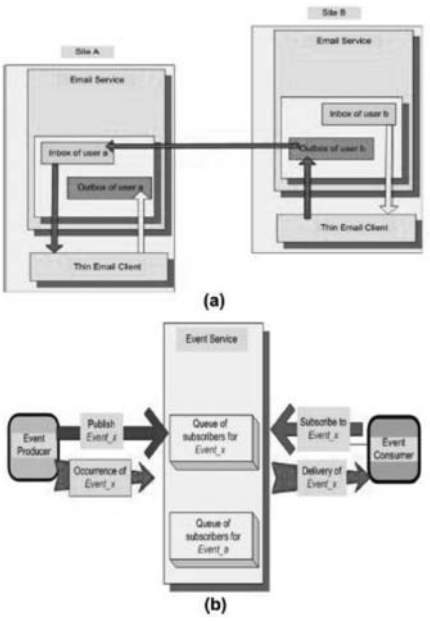

Figure1: (a) E-mail service (b) Event services

8 Analysis and Discussion

In [5] email security summarized in the following issues:

Demanding passwords means strongest password, stopping data leakage with content filtering, stopping spam before it really stinks, stopping breaches with content filtering, making malware go away, blocking breaches, consider compliance, training and best practices, fighting phishing, implementing defense in depth.

Some researchers depend on checking the improvement of keeping the system out of any unwanted effectiveness as the following:

In 2009 Spam volumes increased affectedly, to over 200 billion every day, the Radicati Group guesses that by 2013 more than 80% of global email traffic will consist of spam mail. So this means that employees must offer part of their work time to remove spam's effects, which results in a decrease in productivity and an increase in frustration. The main cost of spam represented in loss of productivity, especially so many spam mails are received every day. Also the cost of bandwidth wasted by spam, additionally storage and network infrastructure is costs too. Furthermore, with the influx of spam and its deletion, an important message could accidentally be trashed along with the unsolicited mail in the rush to clear one's inbox of junk mail [8].

If an employee receives just five spam mails a day and spends 30 seconds on each as Ferris Research calculated, this mean he will waste 15 hours a year on junk mail, when multiply that by the hourly rate of every employee in the company and we will get a very conventional idea of the cost of spam to the organization. Ferris Research, a San Francisco-based IT market research firm, guesses that spam cost a total of $130 billion worldwide in 2009. Other researchers depend on country distribution of the IP addresses that harvested the email addresses that disseminated to know the effect of spam spreading to limit the affected areas as in [7].

Sometimes in order to save time it is necessary to put a break to bandwidth, spam and money. The step to succeed this aim is to direct the network users to save their email address private. However, apart from implementing common sense, an effective server level anti-spam tool needed to be deployed.

There are new attacks called ThingBots, they used novel devices to launch botnet attacks. A single botnet may take control of over 100,000 of these small devices to spread its attack. Actually over than 750,000 spams and phishing messages being sent out in only two short weeks.

Some researchers made statistics depending on geographical area, others may add some additional steps to the encrypted security algorithm, it means the e-mail security protocols, the question is why we not to work to construct robust standard tools to check the system security performance to make monitoring and

supervising the performance system security more efficient and easier to check and get results. This may need additional constructing and designing capabilities for such a system with powerful programming but finally finding dependent standard tool make the work easier and more efficient.

9 Conclusion: Future Directions and Suggestions

In the following points the necessary conclusions to measure the system security performance to reduce as possible or to remove entirely the hacking and spam effects:

- Expecting issues: We should always put in our consideration issues that may appear.
- Educating technical staff: Those who use the system and applications must be well educated and trained in order to create better defense against any attack appears.
- Previous researchers' results: In most other references some statistics are done to measure or compare/show filters or anti spams effective.
- Depending internal standards: for measuring: It would be better to take some standards represented as criteria depending on one variable or two to check the performance of each filter or anti-spam according to dependent standards system security.
- Constructing Final dependent standard criteria: Constructing total criteria depending on previous step to represent a standard tool to check the system security performance by the system security supervisor.
- Benefits of analyzing final criteria: The last final total criteria can be analyzed to find most important benefits as the following in continuous periods.
- Extracting system weakness points: extracting weakness points in anti-spams and filters in order to improve in future.
- Finding strongest tools: Using the previous step to send an instance reports to the supervisors to depend on strongest tools to improve the existent security system.
- Informing programmers and developers: to help anti-spam and firewall developers and programmers to process the weak points and produce strongest anti-spams and filters.
- Avoiding hackers' new technologies: because always hackers are discovering new technologies to penetrating and devising more efficient tools to avoid their effects and thus note discovery.

Acknowledgments. This paper is written under the supervision of Prof. Maria Nisheva-Pavlova and Prof. Vladimir Dimitrov in the Faculty of Mathematics and Informatics, University of Sofia.

References

[1] Robert Shimonski, Sean-Philip Oriyano: Client-Side Attacks and Defense. Elsevier Store, Waltham USA (2012).
[2] Mukherjee, Ranadeep; Dutta, Ambar: An Improved Certified E-mail Protocol Based on Author Based Selective Receipt. International Journal on Cryptography and Information Security (December 2012).
[3] M.Tariq Banday: Effectiveness and Limitations of Email Security Protocols, International Journal of Distributed and Parallel Systems (May 2011).
[4] GFI White Paper, Hosted or on-premise? Choosing the correct option(s) (2011).
[5] GFI Mail Essentials and GFI Mail Essentials Online, 10 things you must do about email security Disclaimer (2014).
[6] Wikipedia, http://en.wikipedia.org/wiki/Email_spam (last modified on 26 Aug 2014).
[7] Gianluca Stringhini, Oliver Hohlfeldy, Christopher Kruegel, Giovanni Vigna: The Harvester, the Botmaster, and the Spammer: On the Relations between the Different Actors in the Spam Landscape. Kyoto, Japan (2014).
[8] GFI White Paper, How to keep spam off your network: What features to look for in anti-spam Technology, USA Canada (2011).
[9] InfoSec Institute, http://resources.infosecinstitute.com/vlan-hacking/.
[10] Dan C. Marinescu: Cloud Computing Theory and Practice (2013 Elsevier Inc.).
[11] Nilotpal Chakraborty: International Detection System and Intrusion Prevention: A Comparative Study (2 May 2013).
[12] GFI White Paper Why one virus engine is not enough. Canada USA (2011).

The use of the m-banking in the Republic of Macedonia

Marina Blazekovic, Viktorija Stojkovski, Monika Angeloska – Dichovska

Faculty of Administration and Information Systems Management, Bitola, R. Macedonia,
University „St.Kliment Ohridski“ – Bitola,
Bitolska bb,
7000 Bitola, R.of Macedonia,
marina.blazekovic@yahoo.com, stojkovskiviktorija@gmail.com,
angeloska_monika@yahoo.com

Abstract. Operating within the modern environment, personal needs and expectations of customers are changing, they know more, want more and do more. Therefore, banks in the Republic of Macedonia seek to create new opportunities by offering financial products and services that are appropriate to these changes, which are easily accessible by modern distribution channels. In order to be a leader, innovator and dynamic, the bank, except that have to understand, also need to succeed at an early stage to uncover customer needs, in order to timely make the necessary investments and changes in its business model and constantly be a step ahead in the banking sector. Innovative solutions by banks consistently facilitate customers every day. With the opening of i-bank store, all alternative innovative banking services get their own space. I-bank store aims to bring, present and offer i-bank alternative channels including and the m-banking. Using a “smart phone”, only with the m-banking mobile application from the bank, customers can to realize some basic banking transactions (paying bills, tax etc.. Transfer of funds and to have constant insight into their products, accounts, statements, a list of transactions and easier to manage their finances. With their mobile applications they have information for exchange rates, interest rates, products, calculators etc. and it is a great advantage in today’s dynamic economic conditions.

Keywords: bank, banking products, i-bank store, m-banking

1 Introduction

According to the research done by the State statistical office of the Republic of Macedonia in 2013, increased the percentage of internet use by households compared to 2012 for 6.8%, while the whole percentage of Internet users is 65,1%[5]. This research forms the basis for banks to be able to focus on creating new modern banking products and services. This type of service has evolved into mobile banking in the year 2011/2012.

Recent trends in the development of e-banking are mobile banking. This kind of service, primarily intended for all who want to be in step with the latest trends in banking operations and save time and money in pursuit of bank services. It

provides easy and convenient use of banking services anytime and from anyplace with usage of mobile phone.

The latest generation of mobile phones allows connecting to the internet directly from the phone, which means that the user can through it to access the bank and perform banking transactions.

Its biggest advantage is mobility, e.g. services provided by mobile banking is independent of the location of the user and the operator , thereby banking services are available 24 hour, 365 days from any place in the world that is covered with GSM signal. Therefore, transactions are executed in a few minutes without going to the bank, so the user saves time and money.

2 Use of mobile banking in banks in Macedonia

The In the first years of its introduction, to the mobile banking was access from the mobile phone via SMS or through WAP standard. The rapid development of technology has contributed to today use technology that enables 3G/4G mobile internet access much faster.

Because the number of mobile phone users is far greater than the number of Internet users and because of the benefits it offers, predicts that mobile banking in the coming years will experience greater expansion than internet banking.

Characteristic for all banks are traditional products and services they offer, such as loans, credit cards, accounts, etc., but not all banks have available modern banking products and services, specifically, the services offered by mobile banking. In the banking sector in Macedonia, only 4 banks use mobile banking in its operations. The use of mobile banking in the Macedonian banking system is in different ways as follows:

SMS banking. The bank that use mobile banking is still in the beginning of its use, i.e., it uses only **SMS banking**. SMS Banking service is an information service of the bank that allows the holders of credit cards issued by the bank to receive SMS information on the status and changes in accounts of their mobile phones. Bank sends messages for the changes made to the accounts of payment cards per individual transaction at the mobile phone.

Fig. 1. SMS banking services

Advantages of using SMS banking service to individuals:

• monitoring the condition of the card 24 hours a day, 7 days a week;
• pprotection of multiple implementation of the same transaction;
• prevention of misuse of the card;
• Information on the card anywhere in the world;
• additional security when conducting transactions.

User of SMS Banking service can be any person who:
• have a transaction account;
• User of debit and / or credit card;
• owns a mobile phone in network authorized mobile operators in the Republic of Macedonia.

Also, this service can be used by entities. SMS banking services for users of bank cards for entities include the following services:

- information for each transaction of payment card service user with information about the account balance of payment card
- warnings for arrears,
- Other information.

S – Token. The banks used the mobile banking application software through the **"S- Token "**, which serves to identify the user via a mobile phone to log

in and verify the payment system for e-banking "Net Banking" for individuals. The only thing necessary for the use of this service is a one-time installation of the application on your "smartphone" mobile phones. Each further use of the application is not required access to the Internet and not any additional devices, allowing easy, simple and practical use [3].

S-Token is a software application installed on the mobile phone and is used for identification (login) user authentication through Net Banking (signature) transactions in the electronic banking system for individuals. S-Token has applications in electronic banking. The advantage of this application is that it is installed on the mobile device and makes it easy, convenient and affordable to use. On the other hand the client does not need him to carry additional devices such as USB token. Services provided by s - Token:

- Review the status of accounts;
- Review the status of loans and credit cards;
- Taking the statement of account;
- Getting a statement of account by e-mail;
- Payment of all persons;
- Payment of borrowings and credit cards;
- Pay their expenses through appropriate forms.

The banks have applied a free mobile application from the **Play Store market**-enabled **Android phones**, from **Windows mobile store** for your **Windows phone** or **i-store market**-enabled for **i-phone** handsets in order always to stay in the first row [4].

M-click. The banks that use mobile banking use application software **m-click** that are installed and used by mobile phone. With m-click application you can now perform various payments directly from your mobile phone and to perform all banking activities from anywhere and at any time [2].

Opportunities offered by m-click are:
- Ability to pay all your regular expenses
- Ability to various payments to individuals and legal entities in the country
- Ability to pay obligations on credit cards
- Ability to review the exchange rate
- Conversion from one currency to another
- Inspect the condition and circulation of all your accounts at any time

Advantages:
- Independence of work hours of branches

• Saving time and money
• Payment shall be made without the need for software token device
• The entry of orders is possible 24 hours, 7 days a week and with a future date

Costs and fees:

• The cost of electronic orders made through m-click, depending on the type of order is reduced from 50% to 70% of the cost of paper orders made in the branches of the bank.

M-click application can be downloaded from the Apple Store or Google Play.

To become a user of m-click - mobile banking, you need to apply to the bank and get an activation code and instructions to use the application.

MobiPay. MobiPay is also a whole new way of cashless payment by mobile phone directly from your bank account. With MobiPay your phone becomes a virtual wallet, and payment is quick, easy and safe [2].

How MobiPay works

Your mobile number can be connected to several different credit cards from the bank. Payments through MobiPay, have to be done in all stores where a tag MobiPay, is there MobiPay terminal.

Payment process takes several seconds, time for entering your first MobiPay PIN in the MobiPay terminal from your mobile device and dials a number for identification. Then it has to be closer to MobiPay terminal. The funds of the payment are deducted from the payment card account that you choose the payment.

Payments are made with a patented technology that allows virtual payment from any mobile device.

Characteristics of Mobipay are fast, safe and smart payment.

Just one click separates you from the simplest way of payment. MobiPay is a whole new way of cashless payment enabled with mobile phone directly from the bank account.

Just wear your cell phone because with MobiPay your phone becomes a virtual wallet, and you get everything you need on the device that is always with you.

Mobile application MBank

MBank. MBank is a free application for mobile phones and tablets, through

which enabled easy and secure access to a large part of the services for individuals. With its use will have permanent access to your accounts, savings accounts or cards, you can easily and safely transfers of funds and to find out all information related to products and services. All you need is to have a mobile phone or tablet and internet connection.

Blend of simplicity and security. Mobile application mBank use modern solutions that offer the most comfortable while using banking services and the highest level of security for customers. For greater safety of users, the application has incorporated maximum limits for daily and monthly turnover. They themselves can change according to your needs.

Currently, this application is unique in the Republic of Macedonia that offers a special interface adapted to work on tablets. The goal is the optimal way to use the surface of the screen and a simplified navigation and access to services [1]. Opportunities offered by mobile banking and its advantages can be briefly summarized in the following table:

Table 1. Opportunities and advantages of the mobile banking

Opportunities	Advantages
- Review the balance and trade of all accounts in any time - Payment of all regular costs - Payments to individuals and legal entities in the country - Payment of liabilities for credit cards - Convert from one currency to another - Overview of exchange rate - Review the locations of bank branches throughout Macedonia - Review the products offered by the bank - Review the locations of ATMs - Using the calculator	- Saving time and financial means - Independence of work hours of branches - Payment without software device-token - Input orders 24 hours 7 days per week

3 Conclusion

The motto of the use of mobile banking is: Simple, faster, cheaper! The banks will continue to upgrade its services and develop their products in order to implement international trends and meet customer needs. Not all banks have available modern banking products and services, specifically, the services offered by mobile banking. The use of mobile banking in the Macedonian banking system is in

different ways. Mobile banking offers a number of advantages and opportunities. Briefly summarized, the possibilities are: review the balance and trade of all accounts in any time, payment of all regular, payments to individuals and legal entities in the country, payment of liabilities for credit cards, convert from one currency to another, overview of exchange rate, review the locations of bank branches throughout Macedonia, review the products offered by the bank, Review the locations of ATMs. Briefly summarized, the advantages are: saving time and financial means, independence of work hours of branches, payment without software device-token, input orders 24 hours 7 days per week. Advantages and opportunities offered by mobile banking are sufficient factor that answers the questions why is needed its greater application by banks.

References

Websites

[1] Sms banking for individuals and Mobile app, Komercijalna banka AD Skopje, http://www.kb.com.mk/Default.aspx?sel=2830&lang=1&uc=1&par=0, 15.3.2014

[2] NLB Klik, NLB Tutunska banka AD Skopje, http://www.nlbtb.com.mk/Default.aspx?mid=395&lId=1, 10.3.2014

[3] IndividualNet Plus – Stoken, Sparkasse Banka, AD Skopje, http://www.sparkasse.mk/individualnet-plus-stoken.nspx, 10.5.2014

[4] M-banking mobile app, Stopanska banka AD Skopje, http://www.stb.com.mk/ns_article-m-banking.nspx, 12.6.2014

[5] Usage of information-communication technologies in households and by individuals, State statistical office, http://www.stat.gov.mk/PrikaziSoopstenie.aspx?rbrtxt=77, 12.3.2014

Architecting Cloud Super Layer of Open Source Components

Hristo Hristov, Vasil Georgiev

Faculty of Mathematics and Informatics, University of Sofia,
5 James Bourchier Blvd., 1164 Sofia, Bulgaria
ico.dimov@gmail.com, v.georgiev@fmi.uni-sofia.bg

Abstract. The purpose of this article is not to put strong formal definition of cloud computing in abstract science terms, but rather a try to explain, "the cloud" and "it's computing" in understandable and geek-free fashion. The article aims to show/list some key Open Source projects, which may serve as a building blocks for a complex and multi-layered cloud environment. We briefed common ingredient classes of this environment which is suitable for widest range of applications as use search engines, public email services and social networks.

Key words: *Cloud architecture, Layered model, Software components.*

1 Introduction

"Cloud computing" has a different but quite close to each other meanings according to surrounding context it is applied for. In computer science, for example, cloud computing is another term for distributed computing based on a network. The meaning incorporated behind this term involves the ability to run a single program consisting of several concurrent processes on many connected computers at the same time. On other hand, in computer networking, cloud computing refers to a [typically large] number of computers connected through a communication network as Internet for example which either execute common tasks or provide infrastructure to various service providers and their customers.

Although we stated two different definitions for a term "cloud computing", which actually sound very much alike to one another, they are not. The common thing they share – connectivity. In both cases we mean a set of computers connected by communication network. The difference between them is in orchestration. By orchestration we mean the invisible layer of software, which makes this set to look like a team of colleagues working together on the same problem, in case of distributed computing. And like a crowd of commuters, who are sharing a common transportation vehicle and ideally do not notice the presence of the others aboard, in case of computer networking.

Given a group of network-enabled computers, which perform a common task or just exchanging messages, we have a near endless field of network-

based services (i.e. the software layer, which makes each and every one of them available). These services appear to be provided by real hardware machines. In fact they are provided by a virtual hardware, simulated from software, executed on real hardware machine (quite often more than one). And this is another aspect of cloud computing called virtualization. Such virtual servers do not physically exist and are quite easily moved around, without any sensible affection on services provided. Their "hardware parameters" are also quite easily changed and scaled up or down, which gives a very important feature of "the cloud" – its elastic behavior.

Considering listed descriptions, Cloud Computing term has a lot in common with quite different and also quite close aspects of technology and computer science. Let us now go into a little bit more details and list the most common cloud computing models in use today.

The thin layer across all cloud computing descriptions points us several basic building blocks of the cloud computing – computer (or processing) power, computer networks, network enabled services and of course a sort of orchestration software layer, which connects all the blocks together. On the other side of the line are the potential users – so called cloud clients. This gives us basic cloud computing classification models, which are: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS).

Let us, state a few words for each of these classes:

IaaS – is the most basic cloud-computing model. It gives us the very foundation on top of which other models exist. Technology building blocks involved here are: virtual machines, computing servers, storage servers, networks, i.e. the very 1st level of cloud computing ingredients.

PaaS – one level up, gives us a whole computing platform. It typically includes operating system, database, web server and programming language execution environment. Each application developer receives ready-to-use environment, where he or she may deploy a particular software solution. The model eliminates the complexity cost of setup and management of lower building blocks – those on hardware and network layers.

SaaS – the highest and probably most complex level. Here, cloud providers install and operate ready-to-user application software, which utilizes resources from other two models. Quite common for this model is application software spread across many virtual machines (often at different geographical locations) with complex load-balancing and network infrastructure. SaaS is labeled as on-demand software, or quite often "pay-as-you-go" software solution. A typical Customer relationship management (CRM) system, Enterprise resource planning (ERP) system or even Email service fits in this class.

Cloud clients – represent the other side of the coin. Typical web browsers, a smartphone with its (possible hundred) mobile applications or

even a modern Smart TV, are typical cloud clients. All of them comprise a software application layer, which communicates and utilizes one or more cloud-situated services.

2 Layered cloud architecture

The lowest component of cloud computing is the operating system. We need a network operating system enough robust and flexible as foundation. By network operation system we mean not only operating system with a communication stack, but such one specially designed for and merged (and still merging) with the Internet.

Natural choice for such operating system could be the GNU Linux, which kernel celebrated its 21-years anniversary last year. Yes, it is quite mature, quite stable and often quite hostile to newcomers because of its UNIX roots. There are many GNU Linux distributions optimized for various purposes: for tablets and laptops, desktops, enterprise desktops, enterprise servers, security enhanced nodes, multimedia (and real-time in general) servers. The choice of OS makes a responsible dilemma on the very beginning of bounding the cloud infrastructure. That is because Linux as an operating system is quite often associated with steep learning curve and complex, terminal-based command line environment. Yes, it is such an operating system, and no it is not. It can be quite friendly and hospitable. Friendly, because of thousands of individuals contributing to GNU Linux as an open-source initiative and they'll never refuse to advice or even help to newcomer. Hospitable, because of the flavors with nice and catchy graphical user interface, where it is possible to do almost every casual daily task, without even knowing that there is Terminal application. And the chosen is – Ubuntu GNU Linux [1]. As it name suggests, it is made for common users rather than advanced developers. It had started as an OS-for-humans effort with pretty and understandable graphical user interface, small footprint, simple and explanatory user guides. Nowadays, it is becoming a software ecosystem of its own which references to all aspects of cloud computing.

Before going any further in Ubuntu details, let us return to our basic cloud classes and draw a reference between these classes and core GNU Linux (not only Ubuntu) technologies for each of them.

Let's start from the 1st class – IaaS. As its name and description suggest here are the very foundations of cloud computing – and major one of them is virtualization. And how this is achieved inside GNU Linux environment? The best in open-source virtualization - Xen project [2], i.e. #1 open source hypervisor. And KVM project [3], i.e. the Linux kernel based virtual machine. Both of them implement a thin software layer to control, schedule and distribute hardware resources between one or more guest operating systems. They also make isolated

environment for every particular operating system, installed on a given Xen or KVM managed machine. This software layer is called hypervisor. As technologies, which reside on such a close-to-hardware level these core projects are fulfilled with hardware related and operating system terminology and are pretty hard and time-consuming to grasp in. Besides they both are good starting if one really wants to see and even "touch" virtualization of computer and network resources on this very low level.

Making a good impression of IaaS could not be achieved without the major providers. Namely - VMware's ESX/ESXi and Microsoft's Hyper-V. Both are best in class solutions, with nice looking and well-set graphical user interfaces.

The 2nd level of cloud is PaaS, and limiting our thoughts to the basic definition we stated, we can say that the GNU Linux provides all that we need. Considering only the well-known, LAMP stack (Linux, Apache, MySQL and PHP) we have fully functional development platform with ready to use knowledge in terms of code repositories, articles and blogs. And what about the complexity of administration and management then? Well, it is a choice of a flavor. That's why we picked up Ubuntu GNU Linux as a host operating system for our cloud-computing journey – almost every set-up you can try to do there is a matter of a limited set of well-defined commands. And you pick up a command line terminal, if there is no nice GUI tool, which can do a setup for you.

The 3rd and most complex class was SaaS. Concluding from definition, we deal with a complex and possibly large software systems, quite often spread across different logical computing entities (virtual machines). We do not consider knowledge needed to use, administrate or develop particular SaaS software system – a CRM for example. Another important part, besides the particular software system, is the orchestration layer needed in order to make such a large software installation fully functional. Again, two major open source projects reside in this field – the OpenStack project [4] and Apache's CloudStack project [5]. Both are quite mature, both can be used in heterogeneous computing environments and both are open source software. But before starting with building its own cloud, the one should consider investigation of basic cloud building block. OpenStack can be used as a good starting point for that – it is well-documented and quite vital software project.

Remember out choice for network operating system – Ubuntu? We stated that it was choice in terms of easily understandable and human-speaking open-source operating system. Nowadays, Ubuntu is even more. It tries to address all listed aspects of cloud computing environment – starting from the infrastructure (including its orchestration) and reaching the end-user device, namely cloud clients.

These were according to us first and introductory steps to an open-source cloud-computing journey. It may look quite long, but it definitely can be quite interesting. And last, but not the least – this journey definitely has a start, but it certainly has not an end. Cloud technologies are constantly emerging and they will be one of the hottest topics in IT/CS fields in the years to come.

Acknowledgements

This research is supported by the project ДДВУ02-22/20.12.2010 of the National Science Fund.

References

1. Canonical Ltd. Ubuntu now on Google Cloud Platform (http://www.ubuntu.com/)
2. The Xen Project. The Xen Project™ Powers the largest clouds in production (http://www.xenproject.org/)
3. KVM Project. Kernel Based Virtual Machine (http://www.linux-kvm.org/page/Main_Page)
4. Openstack. Open source software for creating private and public clouds.. (http://www.openstack.org/, Oct. 2014)
5. Apache CloudStack. Announcing Apache™ CloudStack™ v4.4.1. Mature, easy-to-deploy Open Source Cloud computing software platform boasts improved efficiency and performance. (http://cloudstack.apache.org/, Oct 23 2014)

Measuring Influence of Genome Annotation Version to Data Analysis Results

Ognyan Kulev

Faculty of Mathematics and Informatics, Sofia University, Bulgaria
okulev@fmi.uni-sofia.bg

Abstract. Genome annotation is representation of the ever enriching knowledge about molecular background. It is used as a foundation in all types of data analyses that lead to medical or biological decisions. Genome data analysis results are highly dependent on genome reference annotation, yet the specific version of genome annotation usually does not get attention when discussing results and comparing them. The study is focused on measuring change of results when using different annotation versions and predicting these changes using machine learning methods. Results presented in the paper assess the correlation between measured changes in annotation and measured changes in data analysis results based on different annotations, and evaluating performance of many machine learning methods for the purpose of choosing the most promising ones for future studies.

Keywords: genome annotation, genome analysis, measurement, annotation changes, machine learning

1 Introduction

Genomic data in bioinformatics consists of raw data received from sequencing machines. No meaningful analysis can be run by using only raw data. For that purpose, genomic annotation is needed to mark and explain the structure and function of parts of genomic data. Improving annotation is crucial for correct interpretation of data and it is an ever going process. While these incremental changes to annotation certainly improve accuracy, research is lacking in predicting how improved annotation would affect data analysis results. Measuring improved annotation has two aspects. One is how much annotation is changed, and the other is how much data results are changed when using different versions of annotation. Quantitative measurement of these changes is of particular interest.

There are many types of data analyses that can be used for measuring. Differential expression analysis of genes is widely used data analysis and is suitable for this purpose. Genes encode proteins and the latter are the building block of any organism. The process of decoding of genes and producing proteins is dependent on many factors and is of great interest to scientists. The quantity of produced proteins is the expression levels of genes. Comparing expression levels in different samples is important data analysis that has many applications. It is

convenient data analysis to be used in comparing annotation differences because the end result can be easily quantified and compared.

Gene annotation in this paper is considered only as multiple regions in DNA that have associated identifier with it. These regions can be spread into several contiguous DNA sequences that are close to each other. There can be different splicings for same gene.

2 The IT Context of Genome Annotation

Genome annotation explains the data model for the DNA of an organism. Genome annotation includes various information concerning the molecular content of a studied organism. An important application of genome annotation is in gene expression studies. For actual measurement of expression level of genes, sequence machine takes biological samples and gives short fragments, called *short reads*, which in informatics context are represented as strings of four-letter alphabet characters. These short reads are mapped onto *reference sequences* which are consensusly-ordered strings that represent whole DNA of a particular species. Using annotation, short reads define different combinations of data structures in the DNA sequences detecting genes. The newly detected genes, annotated and defined by certain traits, represent a discovery of a new knowledge. The mapping process which is illustrated in Fig. 1.

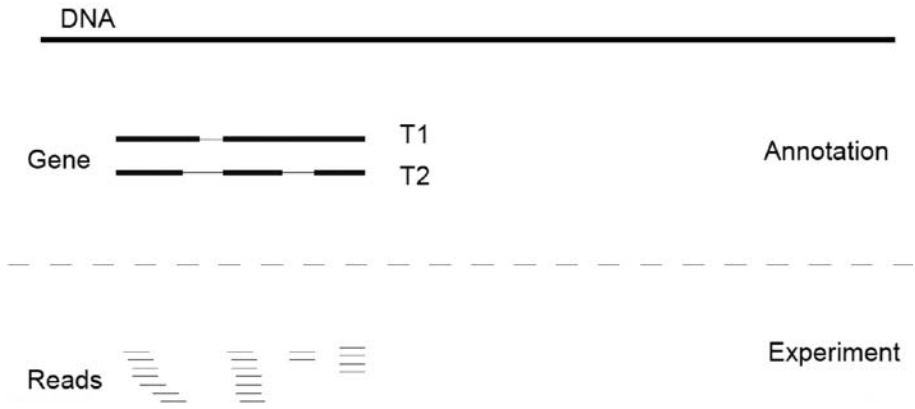

Fig. 1. Mapping of short reads onto reference sequences of gene

Many short reads can be mapped onto either alternative sequence of a gene. They are distributed among the alternative sequences, so that each short read is mapped exactly to one sequence. Other short reads may have only one possible mapping. The number of short reads mapped to specific reference sequence is counted and this is the raw measurement of expression level of gene. These counts are calculated for the samples of the two conditions that are compared.

Differential analysis compares two conditions by comparing these counts and evaluates how much significant is the change and if this significance is above certain threshold, the gene is considered differentially expressed. The final result is boolean value whether gene is differentially expressed.

When two different versions of an annotation are used, final results from differential expression analysis can be compared and difference can be detected. These differences after change in annotation are what we are studying and trying to predict. Machine learning methods are suitable for this task, because there is no algorithm for measuring how changes in annotation affect changes in differential expression. Machine learning would create representation of such algorithm. This study is focused on experimenting with different machine learning algorithms and evaluating their performance. For this purpose, machine learning toolkit with variety of implemented machine learning algorithms is needed.

3 Related works

An increasingly large number of novel genomes are being sequenced and the task of automatic genome annotation has never been more important. The annotation of a single genome is an intensive computational task, having strong biological perspective. The current revolution in sequencing technologies also allows us to obtain a detailed picture of the whole complement of expressed RNA transcripts.

For the purpose of an accurate annotation, it is necessary to have in use methods and software tools for more exact quantification of the expression levels. In some initial analyses was used a simple quantification method [1,2] which counts the number of short reads that map uniquely to each gene, possibly correcting a gene's count by the "mappability" of its reference sequence [3] and its length. The major problems with this type of method are the biased "mappability" and incorrect estimates for the genes with alternative reference sequences. A couple of methods were later developed that addressed the first problem by "rescuing" short reads that mapped to multiple genes ("multireads") [4,5]. Some other methods addressed the second problem, but not the first, by modeling genome data at alternative reference sequence level [6]. Also there were developed a number of methods utilizing different similar statistical approaches methods aiming at more accurate estimation of the abundances at both alternative reference sequences and gene levels [7-13].

4 Methodology of Research

Different organisms can have very different profile of their annotations. Basic properties of annotation profile are number of genes, average number of alternative reference sequences for each gene, average length of gene sequences and others. For the purpose of general measurement of the effect of annotation

changes, four plant organisms are chosen – Brachypodium, Rice, Arabidopsis and Maize. The newest and the oldest publicly available annotations of each of these organisms are used. There are many experiments on these organisms that can be publicly accessed and freely downloaded and analyzed. For each of the four plants, experiments are downloaded and used that compare two conditions that affect many genes. Differential expression analysis is run between the two conditions for each annotation version. For each gene, final result from software is a numerical value that is compared to threshold value to determine if the gene is differentially expressed or not.

One of the two used software packages for differential expression analysis between two conditions is BitSeq [14]. Result from BitSeq is numerical value called Probability of Positive Log Ratio (PPLR). It is calculated as log of ratio of expression in condition 2 over expression in condition 1 or log(Exp_2 / Exp_1). When the ratio is positive, it simply means that the gene expression in condition 2 is higher than the expression in condition 1. Using PPLR values requires setting threshold, and in this paper we used 0.05 as PPLR value threshold. PPLR values are between 0 and 1, and values below 0.05 or above 0.95 signify different expression level between conditions. In the result of differential expression analysis, the outcome used is boolean value whether the gene is expressed differently between the two conditions or not. Having two annotations that are compared, old and new, results in two boolean values for each gene that indicate whether it is differently expressed. Due to difference in annotation, it is expected that some of these pairs of boolean value will be different. Comparing quantitatively these differences and the difference between the two versions of the annotation is the aim of the paper.

The other used software for differential expression analysis is edgeR [15]. Final result for each gene is p-value, as used in statistical hypothesis testing. The used chosen threshold value is also 0.05.

As the first step in this study, biology organisms are chosen. Human genome has the most detailed and precise annotation because of its significance in medicine. This makes it unsuitable for the purpose of the study in this paper because changes between annotation versions are very small. Plant organisms are chosen instead. Compared to mammals, variations in plant genome are much greater and they do not have the convenience to be similar to the well-studied human genome. For plants are chosen and their latin names are *Brachypodium distachyon*, *Oryza sativa* (rice), *Arabidopsis thaliana*, and *Zea mays* (maize). In the rest of the paper, their names are shortened to just single letters: B, O, T, and Z respectively. The following public data of experiments published in European Nucleotide Archive [16] are used: SRP026264, SRP038713, SRP022162, and SRP013564.

Next step was choosing the tools for data analysis. In Bioinformatics, usually custom workflows are created for each case, combining different open source Bioinformatics tools. For mapping short reads to reference, Bowtie2 [bowtie] was

used. For emperical analysis of differential expression, two tools with different algorithms are chosen. EdgeR [15] uses R statistical programming language and is popular amongst bioinformaticians. BitSeq [14] uses bayesian inference and claims to account for both technical uncertainty and intrinsic biological variance in data.

Workflow of data processing required automation and tracking dependencies between intermediate files. While there are specialised bioinformatics workflow systems available, a lot of custom data processing was needed that can be conveniently impelemented in POSIX/Linux shell command language. For this reason, POSIX Makefiles were chosen as workflow language that describes processing dependencies. AWK text processing language was used for light data processing where it is possible. Python scripts were used for data processing that could not be expressed efficiently in POSIX shell commands or AWK. Biopython library was used for input and output of data.

Short reads and annotation are the input of the process. Short reads are mapped onto genome reference as described by annotation and this mapping is stored in BAM files that are used for all analysis. BitSeq consisists of multiple tools that should be called in specific pipeline order for each BAM files. Then intermediate results from all BAM files are used as single input to two more steps. Custom Python script is used to process the final result from BitSeq and make it into table that shows for each transcript is it differentially expressed or not in each annotation, and is there difference between the two differential expression results. The other software, edgeR, requires more preliminary processing and it is achieved with HTSeq [17] package and POSIX pipelines of commands. Simple R scripts execute the recommended sequence of commands for edgeR processing. Another custom Python script was used for transforming the output of edgeR to look like result from BitSeq and then again the Python script to produce the summary table is used. These summary tables along with statistical information are transformed into one ARFF file for each version of the annotation.

WEKA [18] is a collection of implemented machine learning algorithms for data mining tasks. Having many algorithms and filters, it is very good choice for the purposes of our study. For WEKA analysis, various parameters have been considered but ultimately the number of mapped short reads onto each transcript was used as an input. These counts are already computed in the described workflow because they are the input of edgeR. WEKA was used to make models of predicting will there be difference in differential expression between using one version of annotation and another version. These two versions were the newest and the oldest publicly available versions of the annotation. Since the number of transcripts that have such difference is very small compared to all transcripts, without preprocessing all machine learning methods of WEKA predict that there will be no difference and statistically this is correct answer. For this reason, SpreadSubsample filter was used to make equal the number of transcripts with

difference and with no difference. Then Normalize filter was used for the counts so that these numbers are in uniform range from 0 to 1.

WEKA has many classifiers and they are divided into multiple sections. In choosing classifiers to use in this study, maximum coverage of different algorithms was desired and approximately one classifier from section was chosen. From decision tree section, open source implementation of C4.5, named J48, and Random forest ensembling learning method were chosen. DecisionTable majority classifier was chosen from rule learners. Meta classifiers are represented with Bagging using REPTree as base classifier. Instance-based learning with k neighbours (IBk) was used from lazy section. MultilayerPerceptron that implements neural network learner is used as representing functions section in WEKA learners. Lastly, NaiveBayes was used from bayes-based learners.

The last choice that need to be made is what criteria to use to evaluate performance of different machine learning methods. The simplest measurement of performance is the percentage of correctly classified instances. Since this is preliminary study of viability of different methods, it is the only measure that is used to compare results. Ten-fold cross-validation was used when measuring classified instances. Since two datasets are used, from BitSeq and edgeR, this study uses one dataset for training and the other for testing. This way each combination of organism and machine learning algorithm is represented in the results as two percentage numbers of correctly classified instances for BitSeq and edgeR when using ten-fold cross-validation, and another two percentage numbers for BitSeq and edgeR but using edgeR and BitSeq respectively as test set.

5 Results and Discussion

In Table 1 are shown the results from the workflow described in the Methodology of Research section of the study.

Table 1. Summary of results of differential expression analysis.

		Total transcripts	Any DE	Any DE, eq	Any DE, eq, %	Diff DE	% of any DE	Diff DE, eq	% of diff DE
B	BitSeq	25195	2684	2681	99,89%	1458	54,32%	1456	99,86%
	edgeR	25195	6980	6969	99,84%	553	7,92%	551	99,64%
O	BitSeq	59950	1397	232	16,61%	1176	84,18%	151	12,84%
	edgeR	59950	2046	371	18,13%	524	25,61%	64	12,21%
T	BitSeq	31304	19529	9394	48,10%	1083	5,55%	468	43,21%
	edgeR	31304	16767	7844	46,78%	2102	12,54%	772	36,73%
Z	BitSeq	122081	1390	1189	85,54%	902	64,89%	752	83,37%
	edgeR	122081	5356	4383	81,83%	1130	21,10%	440	38,94%

Four organisms codenamed earlier as B, O, T and Z and the two programs BitSeq and edgeR are heading of each row. The first column is the total number of transcripts that were present in the studied biological samples. The second column is number of transcripts that are differentially expressed ("DE") using either version of annotation. The third column is like the previous one but counting only these transcripts that didn't change between the two versions of annotation ("eq" is short for "equal"). The fourth column is the percentage of these filtered transcripts from the unfiltered transcripts with differential expression (the second column). The high percentage in all organisms can be surprising at first. It shows that many transcripts are not changed at all between versions and yet in one version there is differential expression and in the other there isn't. One reason is that new transcripts and removed transcripts usually are similar to other transcripts. When new transcript similar to another is added, short reads that were mapped to unchanged transcript are now distributed between the new and old transcripts. This changes number of short reads mapped to both transcripts and leads to different result in differential expression analysis. Another reason for this is that both BitSeq and edgeR implement complex statistical methods that take into account the statistics of the whole biological sample and not just individual transcripts.

The fifth column is the number of transcript that have one differential expression result in one version and opposite result in the other version ("diff"). This classification is what the paper studies and this is the predicted value by WEKA machine learning algorithms.The sixth column is what percentage of transcripts with different differential expression between version compared to transcripts that have differential expression in at least one version (the second column). The seventh column is number of these transcripts that are not changed between versions of annotation and the eighth column is percentage of these filtered transcripts compared to unfiltered. The percentage is very close to the percentage of filtered to unfiltered transcripts that have differential expression in any version. It can be concluded from this comparison that there is no correlation between equality of transcripts in different versions and equality in differential expression analysis result. This suggestion is confirmed in WEKA analysis not presented in the paper. It was observed that all analysis prediction are worse after adding statistic information about transcript differences between different versions of annotation.

Performance comparison of the chosen WEKA machine learning algorithms for prediction of the influence of differential expression to annotation accuracy is given in Table 2.

Table 2. Summary of the WEKA machine learning methods applied to differential expression analyses

Organism:		B		O		T		Z	
Test set →	Training set	BitSeq	EdgeR	BitSeq	EdgeR	BitSeq	EdgeR	BitSeq	EdgeR
J48 (trees)	BitSeq	76,47%	53,29%	82,87%	59,61%	62,88%	57,34%	74,50%	**76,00%**
	edgeR	55,43%	60,58%	69,75%	85,21%	53,62%	87,92%	**83,98%**	89,03%
RandomForest (trees)	BitSeq	78,84%	51,51%	83,55%	65,35%	**72,76%**	**59,28%**	**75,22%**	71,73%
	edgeR	55,43%	**67,81%**	**76,15%**	**90,94%**	59,92%	**90,56%**	83,54%	90,58%
DecisionTable (rules)	BitSeq	73,94%	**63,82%**	77,51%	**71,98%**	67,50%	53,92%	74,33%	75,67%
	edgeR	54,43%	60,58%	74,14%	76,05%	61,01%	82,97%	83,76%	86,81%
Bagging (meta)	BitSeq	**79,46%**	52,64%	**84,01%**	67,52%	72,25%	58,40%	74,39%	72,06%
	edgeR	**57,23%**	66,09%	69,94%	90,84%	**64,82%**	88,96%	83,81%	89,91%
IBk (lazy)	BitSeq	75,69%	50,27%	81,59%	63,27%	60,02%	57,29%	75,06%	64,41%
	edgeR	53,44%	64,47%	65,74%	88,07%	61,08%	83,80%	75,66%	**91,77%**
Multilayer-Perceptron (functions)	BitSeq	63,31%	50,00%	68,71%	66,28%	50,60%	47,88%	63,58%	68,68%
	edgeR	48,64%	50,18%	74,43%	69,27%	50,45%	55,59%	65,18%	70,53%
NaiveBayes (bayes)	BitSeq	56,10%	44,68%	58,76%	52,42%	52,35%	50,69%	67,24%	62,53%
	edgeR	48,28%	50,99%	65,27%	51,72%	50,45%	51,71%	61,68%	57,88%

In Table 2, each row is from using training set from BitSeq or edgeR, while columns with heading BitSeq and edgeR denote the test set used. Same dateset for training and testing means that ten-fold cross-validation is used for measuring. Percentage is the ratio of correctly classified instances. For each organism and combination of training and test set, maximum ratio of correctly classified instances from all machine learning methods is visually marked with bold typeface. Conclusion about the performance of different machine learning algorithms will be based mostly on these marked percentages.

Marked cells with the best performing results show that NaiveBayes and MultilayerPerceptron didn't perform best in any situation, especially NaiveBayes. J48 is best performing in organism Z for the cases when validation is user test from other differential expression analysis. What can be observed about organism Z is that all machine learning algorithms perform very well compared to other organisms. Conclusion can be made that the chosen input and predicted data are well suited for predicting for this specific organism. J48 and IBk successes in this case and no best performance in other organisms means that they can be removed from future consideration.

The best performing algorithm according to this table is RandomForest. Comparing performance of RandomForest and the second best algorithm, Bagging, reveals that numbers in all situations are close to each other. DecisionTable is the third best algorithm by this criteria. If one algorithm need to be chosen for future

consideration, it could be either RandomForest or Bagging, and if possible, both should be considered.

In interpreting the data it is important to remember that predicted value is one boolean that signifies whether or not there is difference in differential expression when using old and new annotation. Due to SpreadSubsample filter, training set has equal number of samples with the two different values. If test set is fixed predicted value, result would be 50% correctly classified instances. Result would be similar if predicted values are random with probability 0.5. Looking back at table 2, results from NaiveBayes are almost always around 50%. This means that this algorithm is no better than random values for these particular training sets. The chosen two best algorithms give results in approximate range 60-80%.

6 Conclusion and Future Plans

Numerous machine learning algorithms were studied in different conditions for prediction of annotation influence and two of them – RandomForrest and Bagging – in most cases perform much better than the rest. Correctly predicted percentage, chosen as a measure for performance, needs to be consistently very high, thus the results of these two best algorithms across multiple organisms are promising.

Further development of the study aimed at expanding of the input data in order to include the changes in annotation and not just the reflection of these changes into counts of mapped short reads.

All the machine learning algorithms were run with default parameters and changing parameter values may improve the results also.

Evaluation process needs to be automated as is done with the workflow for differential analyses.

All these future directions could make the study more complete.

Acknowledgements. The presented work has been partially funded by the Sofia University SRF within the "New approaches and information technologies for building special-purpose knowledge based systems" Project, Contract No. 044/2014, and partially funded by European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

7 References

1. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science 2008, 320(5881):1344-1349.
2. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Research 2008, 18(9):1509-17.
3. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M: Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques 2008, 45:81-94.
4. Faulkner GJ, Forrest ARR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, Hume DA, Grimmond SM: A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics 2008, 91(3):281-8.
5. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 2008, 5(7):621-8.
6. Jiang H, Wong WH: Statistical inferences for isoform expression in RNASeq. Bioinformatics 2009, 25(8):1026-1032.
7. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 2010, 26(4):493-500.
8. Katz Y, Wang ET, Airoldi EM, Burge CB: Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods 2010, 7(12):1009-15.
9. Nicolae M, Mangul S, Măndoiu I, Zelikovsky A: Estimation of alternative splicing isoform frequencies from RNA-Seq data. In Algorithms in Bioinformatics, Lecture Notes in Computer Science. Edited by: Moulton V, Singh M. Liverpool, UK: Springer Berlin/Heidelberg; 2010:202-214.
10. Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, van Baren M, Salzberg S, Wold B, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 2010, 28(5):511-515.
11. Feng J, Li W, Jiang T: Inference of isoforms from short sequence reads. Journal of Computational Biology 2011, 18(3):305-21.
12. Paşaniuc B, Zaitlen N, Halperin E: Accurate Estimation of Expression Levels of Homologous Genes in RNA-seq Experiments. Journal of Computational Biology 2011, 18(3):459-68.
13. Richard H, Schulz MH, Sultan M, Nürnberger A, Schrinner S, Balzereit D, Dagand E, Rasche A, Lehrach H, Vingron M, Haas SA, Yaspo ML: Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. Nucleic Acids Research 2010, 38(10):e112..
14. Glaus, Peter, Honkela, Antti, Rattray and Magnus (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics, 28(13), pp. 1721-1728.
15. Robinson, MD, and Smyth, GK (2008). Small sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9, 321-332.
16. Leinonen, R et al. The European Nucleotide Archive. Nucl. Acids Res. (2011) 39 (suppl 1): D28-D31.
17. Simon Anders, Paul Theodor Pyl, Wolfgang Huber. HTSeq - A Python framework to work with high-throughput sequencing data. bioRxiv preprint (2014)
18. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Business Processes in Grid and Cloud

Radoslava Hristova, Vladimir Dimitrov

Faculty of Mathematics and Informatics, University of Sofia "St. Kliment Ohridski",
James **Bourchier 5**, 1164 Sofia, Bulgaria

radoslava@fmi.uni-sofia.bg, cht@fmi.uni-sofia.bg

Abstract. Basic concept in the cloud is the service. In the last three years a special instance of the SaaS model is a subject of discussion, namely the provision of business process as a service (BPaaS). The idea behind BPaaS is to provide entire end-to-end business solutions as cloud services. The cloud computing has evolved through a number of phases one of which is Grid computing. And regardless of the differences between the cloud and the scientific Grid, the idea of business process management is still the same. In this article we make a brief overview on the problem.

1 Introduction

One of the definitions given for cloud computing defines the cloud as "a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service by using Internet technologies". This definition is given by Gartner [1] and clearly shows that the basic concept in the cloud is the service. The cloud model is composed of three service models [2]: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). In the last three years a special instance of the SaaS model is a subject of discussion, namely the provision of business process as a service (BPaaS). The idea behind BPaas is to provide the entire end-to-end business solutions as cloud services. In [3] the author define BPaaS as "a special SaaS provision model in which enterprise cloud offerings provide methods for the modeling, utilization, customization, and execution of business processes". The author also notes that possible realization of this service model is implementation of the service-oriented architecture (SOA) and WS-BPEL for services' orchestration.

The concept for the service and service composition into business processes is well known in SOA. The service in the context of SOA is function or operation accessible through the network, which has well defined interface, functionality with endpoint for access. Because of the model advantages, SOA has become the standard for enterprise applications development and integration. As regards the term business process, it has very broad meaning. In the context of SOA, the business process is a composite Web service written in WS-BPEL and


V. Dimitrov, V. Georgiev (Editors): ISGT'2014. ISSN 1314-4855
Proceedings of the 8th International Conference on
Information Systems and Grid Technologies, Sofia, 30–31. May, 2014.

executed by a specialized Web service, called "orchestrator". The most popular implementation of SOA is via Web services.

It is very important, business organization to be provided with IT solutions, which to support the business processes from end-to-end. It includes not only the modeling of the business process, but development, deployment, execution and monitoring of the process. It should be taken into account, that a business process can include tasks from different applications, message exchange, integration of different services, even human tasks. Considering the fact that business operations are often changed and complicated, which suggests fast adaptation to the new business requirements. Therefore, the business process management is essential for every business organization. Proper business process management in a given business organization can lead to high efficiency, lower operating costs, reduced waste, and proper utilization of human resources.

Service-oriented architecture has proved its efficiency as a solution for integration and business process management. SOA provides agile environment in which the logic of automation can be separate into services. The services can be composed into business processes. The services are interoperable, loosely-coupled, reusable and discoverable. These features make SOA environment flexible and adaptive to changes. In the context of SOA, business process management is covered by four phases: modeling, assembling, deploying and managing of the business processes. These phases are also known as service-oriented architecture lifecycle [4].

Essential characteristic of the cloud is access on demand and broad network access to the heterogeneous platforms including mobile phones and tablets through standard mechanisms. The cloud technologies provide availability to the users to access heavy applications with respect to the resource consumption through lightweight handheld devices. By moving of computing power and information from desktop computers and portable devices into computational and data centers, the cloud provides a new, flexible approach for doing business.

The cloud computing has evolved through a number of phases one of which is Grid computing. And regardless of the differences between the cloud and the scientific Grid, the idea of business process management is still the same. In this article we make a brief overview on the problem.

2 Business processes and the scientific Grids

The computational Grid is defined in [5] as "an independent, comprehensive, overall hardware and software infrastructure, which offers inexpensive access to high volume of computational resources". In [5], the authors also define major applications classes for computational Grid, namely distributed supercomputing, high-throughput computing, on-demand computing, data-intensive computing and

collaborative computing. Based on the concepts presented in [5], the foundations of a new scientific Grid infrastructure were placed in 2001. Currently, the name of the infrastructure is the European Grid Infrastructure (EGI) [6].

Characteristics of the scientific Grid are resource sharing, secure data access and efficient and balanced use of computational resources. The scientific Grid coordinates use of distributed resources, shared by different institutes, computational centers and organizations. This way of usage of computers, networks and devices provides great benefits for every scientist who needs computational power or storage resource for solving scientific problems. The computational resources in the scientific Grid are dedicated to the infrastructure. They cannot be used independently, outside of the Grid. The access and usage of the scientific Grid is free for the end user.

The idea for the open standards also has been laid in the concepts of Grid. Interoperability between different Grids was achievable only by the adoption of open standards for Grid development. The standardization encourages industry to invest in developing commercial Grid services and infrastructure. The standards in Grid are developed by Globul Grid Forum (GGF). GGF intends to define Grid specifications that can become broadly accepted standards for the international society to exchange ideas, experience and best practices. Open Grid Services Architecture (OGSA) is the GGF's solution for information and resource sharing among organizations, which utilize products from different vendors. The Open Grid Services Architecture (OGSA) defines the term Grid service as a Web service with an extended interface, which to support life cycle status and asynchronous events. Basic advantage of OGSA is presentation of all resources in the Grid as Grid services. This approach simplifies the presentation of the resources and allows consistent access to resources located on heterogeneous platforms. This means Grid based on Web services which is OGSA-compliant is and SOA-compliant.

In [7] the OGSA standard is supplemented with detailed description of basic services and the requirements they have to meet. The architecture defines security services, services for data management, and others. However, the specification presented in [7] does not give clarity on the issues related to service composition into business processes and their implementation. The software implementation of OGSA is realized as the product Globus Toolkit. We can say that OGSA introduces a concept for applying service-oriented architecture in Grid and currently it remains the only standard for service-oriented Grid, which exists. Unfortunately, this standard failed to enforce in the development of scientific Grid environments.

EGI is infrastructure which gradually evolved as a result of many projects. In the different projects the development of the EGI's Grid middleware are followed different software approaches. In the very beginning there were attempts to apply

the standards introduced by GGF. Subsequently in the next Grid middleware version the direction has been directed to the service-oriented architecture. Some service-oriented approaches for gLite Grid environments and EMI were considered. The architecture of gLite and its services was presented in [8] as services which follow SOA. This was pointed by the author as feature which "will facilitate interoperability among Grid services and allow easier compliance with upcoming standards, such as OGSA". All these ideas were laid down into architecture of the middleware. Unfortunately, they were partially accomplished and currently g-Lite middleware is far away from OGSA-compliant. The current Grid middleware - EMI distance itself from the idea of service-orientation.

However, Grid computing is intended to be an environment for business processes and even more, they to be implemented as Web services. In [9] the author discusses the important aspects of service-oriented Grid and underlines the lack of widely accepted mechanisms for orchestration of business processes and their mediation and monitoring. EGI is not an exception. Basic indicators for that are:

- Lack of service registry or support of service registry (UDDI is not supported);
- Lack of discovery services;
- Lack of service for composition (BPEL is not supported);
- Lack of well-defined descriptions of Web services (WSDL is not fully supported);

In [10] the authors analyze existing tools for service compositions in EGI. Some of the tools provide partial solutions to the problem, focusing on the definition and implementation of business processes in EGI, without engaging with the definition of service compositions. These solutions are not based on service-oriented approach and therefore do not support Web services. Other solutions provide the ability to automate part of the processes in order to be executed in Grid. For implementation of the process they rely on the legacy services of the EGI middleware. The management of a business process, however, involves not only definition and implementation of the process, but and the ability of monitoring and optimization of the business processes. The business processes may include human tasks they also may have rules. These features are not provided and therefore are not realized in the analyzed tools.

The information system of EGI contains information on all available services in the Grid infrastructure. Unfortunately, only few of them followed the principles of service-oriented architecture. Thus, some of the services in the Grid infrastructure have WSDL descriptions with information for the endpoint of the service, but for the most of them such information is not available. The business process management in the EGI implies the existence of a layer of Web services between the business processes environment and the Grid infrastructure. Such layer is an

essential part for the process of the service composition. Nevertheless, currently in the EGI grid infrastructure such layer is not provided. These considerations suggest the implementation of additional modules and functionality. Example solution of the problem is presented in [11] and shown on Figure 1.

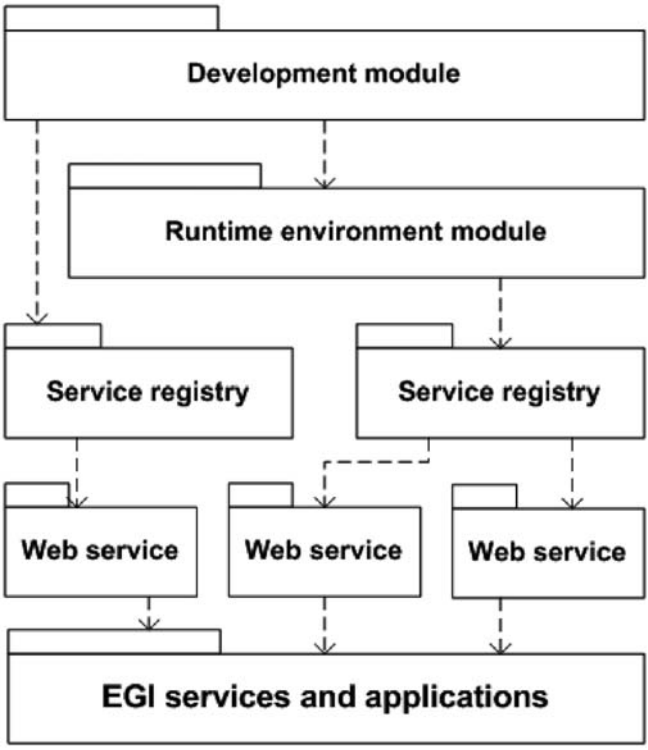

Figure 1. SOA-based platform for the EGI

The presented SOA-based platform consists of three basic modules: Development module, Runtime environment module and Service registry module. As a representation of a model of the platform for the Grid, the Runtime environment module has to be a part from the grid infrastructure. The Development module provides functionality for business process design, development and deployment. The Service registry module is a registry for the developed web services and business processes. The Runtime environment module includes service for business process execution, service for business rules management, service for business process monitoring and service for human task management. All four services use common infrastructure for message exchange – the Enterprise Service Bus (ESB).

3 Business processes and the cloud

However, the direction of EGI is moving away from OGSA, the service-oriented architecture and the business process management. EGI establishes new software integration, based on cloud computing or more precisely on a locally deployed IaaS Clouds [12]. Thus infrastructure will provide generic, consistent and flexible access to EGI resources. The new Core Infrastructure Platform is organized in two distinct platforms: The EGI Core Infrastructure Platform, which will include only services that are necessary for operational federated, distributed computing infrastructure and the EGI Cloud Infrastructure Platform, which will be deployed on the top of the EGI Core Infrastructure Platform. The new cloud infrastructure will provide new federation and distribution services directed to the cloud computing and supports new research communities who wish to deploy their own Virtual Research Environment to the resources that they are able to access.

The open questions that remain in front the scientific grid (cloud) are still the same – usage of open standards and SOA support. As it seems, the business process management, will not be covered and in the newly proposed EGI Cloud Infrastructure Platform.

As a difference of the scientific grid the situation in commercial clouds is not the same. There are several cloud solutions which provide BPaaS. Two of them are the BPM platform of Fujitsu [13] and Appian BPM Suite [14]. Both of them are flexible platforms which are SOA-compliant.

The BPM platform of Fujitsu is available as a cloud service or as a standalone product which can be deployed. Cloud users are provided by PaaS and SaaS based cloud solution where all of the phases in the BPM lifecycle are fully supported by the platform.

The Appian BPM Suite provides web-based drag-and-drop design tool based on the Business Process Model and Notation (BPMN) standard, built on a scalable Service Oriented Architecture (SOA). The Appian BPM Platform is exposed as a Service in the Cloud. It supports modeling of business processes, business rules, managing of enterprise content and related with them artifacts.

4 Conclusions

Regardless of the differences between the scientific Grid and the cloud, the experience shows us that crucial moment in the enactment of one distributed infrastructure is the ability to manage business processes. For scientific grid and for the cloud the solution is the effective realization of the business processes in the infrastructure, based on service-oriented architecture (SOA) and WS-BPEL for orchestration of services.

Acknowledgments. The work reported in this paper is supported by the project Development of Grid Technologies at JINR and SU "St. Kliment Ohridski" - Information, Computer and Network Support of JINR activities", № 05-6-1048-2003/2013 and is partially supported by Sofia University "St. Kliment Ohridski" SRF under Contract 05/2014.

References

1. 2010 Cloud Computing, David W. Cearley, 2010; http://www.gartner.com/it/initiatives/pdf/KeyInitiativeOverview_CloudComputing.pdf
2. P. Mell and T. Grance. "The NIST Definition of Cloud Computing", National Institute of Standards and Technology, September, 2011
3. R. Accorsi. "Business Process as a Service: Chances for Remote Auditing", COMPSAC Workshops, 2011, pp. 398-403
4. R. High, S. Kinder and S. Graham. "IBM's SOA Foundation: An Architectural Introduction and Overview", November, 2005
5. C. Kesselman and I. Foster. "The Grid: Blueprints for a new Computing Infrastructure", Morgan Kaufmann Publishers, 1998
6. European Grid Infrastructure (EGI), http://www.egi.eu/
7. I. Foster, et al. "The Open Grid Service Architecture, Version 1.5", 2006
8. E. Laure, et al. Programming the Grid with gLite, Computational Methods in Science and Technology 12(1), 2006, pp. 33-45
9. V. Dimitrov. "Development of applications with service-oriented architecture for Grid", ACM New York, Proceedings of the 9th International Conference on Computer Systems and Technologies, Article No.14, 2008
10. R. Goranova. "Service composition tools in g-Lite", Conference Proceedings of the 5th International Conference ISGT, 2011, pp. 228-235
11. R. Goranova. "Architecture of a SOA-based BPM Platform for EGI", Conference Proceedings of the 5th International Conference "Distributed Computing and Grid-technologies in Science and Education", 2012, pp.138-143
12. EGI Platform Roadmap, https://documents.egi.eu/document/1624.
13. BPM platform of Fujitsu, https://www.interstagebpm.com/
14. Appian BPM Suite, http://www.bptrends.com/publicationfiles/07-09-Appian%20BPMSuite%20Ver.%205.7%20_3_-Malcolm1.pdf

Business Process Model Based on Business Rules

Evgeniy Krastev [1] Maria Semerdjieva [2]

[1] Faculty of Mathematics and Informatics, St. Kl. Ohridski University of Sofia, 5 James Bourchier Blvd., 1164 Sofia, Bulgaria, [2] Faculty of Mathematics and Informatics, St. Kl. Ohridski University of Sofia, 5 James Bourchier Blvd., 1164 Sofia, Bulgaria
1 eck@fmi.uni-sofia.bg 2mtutanova@fmi.uni-sofia.bg

Abstract. Representational analysis of procedural and declarative representation of business process modeling techniques is one of the most interesting research directions. Previous work has identified a lack of process modeling language capabilities to model business rules. Thus it has been established that the procedural representation doesn't provide explicit information about important business concerns governing the work in an organization. The purpose of this paper is to examine the relations of business process modeling with business rules. The objective is to provide a declarative representation of a business process model providing explicit information about business concerns governing all the possible instances of the business process model execution. The main idea is to separate the business concerns in the declarative representation of a business process model and those in the execution of a business process. For this purpose we define a trajectory in the space of business activities, where the business concerns in the declarative representation define the valid segments for movement along that trajectory. Accordingly, the valid movements along this trajectory define the set of business process execution instances. The business rule vocabulary of the EM-BrA²CE Framework is being used in a case study to illustrate the here proposed approach for declarative representation of business process models.

Keywords: Business Process Modeling, Business Rules, BPMN, Model representation

1 Introduction

In view of the growing complexity of today's information systems the representation of business process modeling with business rules is gaining a lot of attention by both academic and industry communities. Currently it is common practice for business process models like Business Process Model and Notation (BPMN) [1-2] and Event Driven Process Chain (EPC) [3] to embed implicitly business rules in their business logic. On the other side, business rules cannot exist outside the context of a business process. The synergies and overlapping representational capabilities between modeling languages for business process and modeling languages for business rule are subject to extensive research

efforts [4- 5], where the basis for comparing different modelling techniques has been the Bunge-Wand-Weber representation model. Thus, the addition of Simple Rule Markup Language (SRML) to the domain of BPMN may further enrich this construct-rich process modeling language [6]. In short, these research results provide evidence in support of the hypothesis that the synergy of rule specification languages and modeling languages for business processes offers the highest representation in all four clusters, Thing, State, Event, and System of the Bunge-Wand-Weber representation model [7-8].

The principle of Rule independence [9- 10] introduced in the Semantics of Business Vocabulary and Business Rules (SBVR) [11] standard restricts the business process model of dealing explicitly with typical business concerns as cycle time characteristics, costs, constraints benefits and business goals of the process. This has led to the development of declarative representation of process models as a counterpart to the procedural representation of process models satisfying the Rule independence principle. The EM-BrA²CE Framework[12] is one of the typical examples for the implementation of declarative representation of business process models. The declarative representation is also referred to as rule- based representation of a process model because the logic of its control flow, data flow and resource allocation is declaratively expressed by means of business rules. This convergence of business process modeling and business rules appears to be the next OMG challenge because it raises several important problems to resolve. For instance, “business process” and “event” are not explicitly defined in SBVR.

The goal of this research is to present an object-oriented approach for declarative rule- based representation of a business process model. The EM-BrA²CE Framework is employed for this purpose. The thus proposed hierarchical structure of a business process complies with the SVBR specification of OMG and allows taking in consideration the business concerns of a business model in explicit form. Moreover, it allows adapting the XML patterns of the SVBR specification in transforming the declarative rule-based representation of the business process in XML format.

2 Problem statement

The purpose of SBVR is to provide semantics and formal representations of controlled structured natural language. The goal of SBVR is to define formalism for natural language presentation of business vocabulary and business rules. An advantage of SBVR is that they are understandable by people without IT skills. The SBVR specification follows the Business rule standards and formal representation and it is based on first order logic, alethic modal logic and deontic logic. Additionally, SBVR adopts a theoretical model [9] for semantic

formulations in terms of structured English patterns represented originally by XML Metadata Interchange format (XMI). It's difficult to employ the current SBVR XMI model in software tools. The XMI representation is suitable merely to represent single concepts, terms and fact types of this model in an isolated way; however, it cannot provide the design of a uniform structure of all the categories in SBVR with all the relationships among them. In an attempt to partially resolve this problem the latest edition of the SBVR standard [13] allows mapping to Meta Object Facility (MOF) and provides UML to MOF mappings. Therefore, the complexity of SBVR and the lack of standard specifications for transforming XMI to an object oriented data model suitable for computer processing makes it rather difficult to use it in software tools.

EM-BrA²CE Framework, for instance, is used to express the business process vocabulary and the business rules related to business process definition only in Structured English, which just partially can be represented in XMI constructs borrowed from the SBVR standard. SBVR gives only a **partial** hierarchical and inheritance representation between these categories. XMI representation of **inheritance** and **hierarchical** relationships between the components of the SBVR vocabulary is achieved using MOF. The SBVR Vocabulary is mapped to MOF elements that make up the SBVR XMI Metamodel. Hereby we distinguish SBVR and MOF both in terms of syntax and semantics. For instance, a MOF class is not the same thing as a SBVR concept and there is no semantic equivalence between MOF and SBVR[9]. SBVR offers MOF Elements of the SBVR XMI Metamodel with purpose of description of the mapping of the SBVR Vocabulary into a MOF-based meta model thus allowing the transformation in terms of UML and object orientation representation of the data model.

The main problem we consider in this paper is to propose a mapping of the SBVR business process vocabulary into a hierarchical structure employing inheritance relationships between the components. At the same time the proposed model should keep the component properties described in their XMI representation in the SBVR standard.

3 Representation of a business process vocabulary with OWL

SBVR does not have a vocabulary defined for expressing business process related concepts. EM-BrA2CE (Enterprise modeling using Business rules, Agents, Activities, Concepts and Events) Framework extends SBVR vocabulary with process-related concepts. EM-BrA2CE adds concepts, fact types and business rules and defines business process vocabulary which can be used in the context of rule based declarative process modeling according SBVR specification.

SBVR was proposed to serve as "transformation of the meanings of concepts and business rules as expressed by humans into forms that are suitable to be

processed by tools, and vice versa". OWL 2 [13] is developed to be one of that forms. OWL correspondences are proposed in SBVR specification. Currently, there are many studies how formal logic on SBVR can be transformed in OWL, but detailed study how this transformation can be done is still missing.

Our objective is to find a uniform common XML based hierarchical representation of the business process categories. For this purpose, first we identify XMI patterns in SBVR related to describing terms, concepts and fact types and next these patterns are being transformed in OWL. Fact types with role bindings can be expressed with RDF, which provides XML hierarchical representation of vocabulary.

According to SBVR we should define terms, concepts and fact types to enable business process modeling. Following EM-BrA2CE at level of vocabulary we will define the following concepts:

- Activity
- Agent
- Role
- State

The following fact types are related to the above concepts according EM-BrA2CE Vocabulary:

- Agent pertains to agent
- Agent can have Role
- Role can perform Activity type
- Activity has state
- Activity can consist of Activity
- Activity has coordinator Agent
- Activity has performer Agent
- Activity can perform assert a Business fact type
- Activity can perform retract a Business fact type

The relationships between these concepts can be represented in the UML class diagram displayed on Fig 1. The Activity concept entity is associated to the Role, Agent, State and Business fact type concept entities through aggregation relationships. An Activity concept entity can be composed of other Activity concepts. This object oriented model allows to use OWL and RDFS to transform the XMI pattern representation of the business Vocabulary into o a hierarchical object oriented data model.

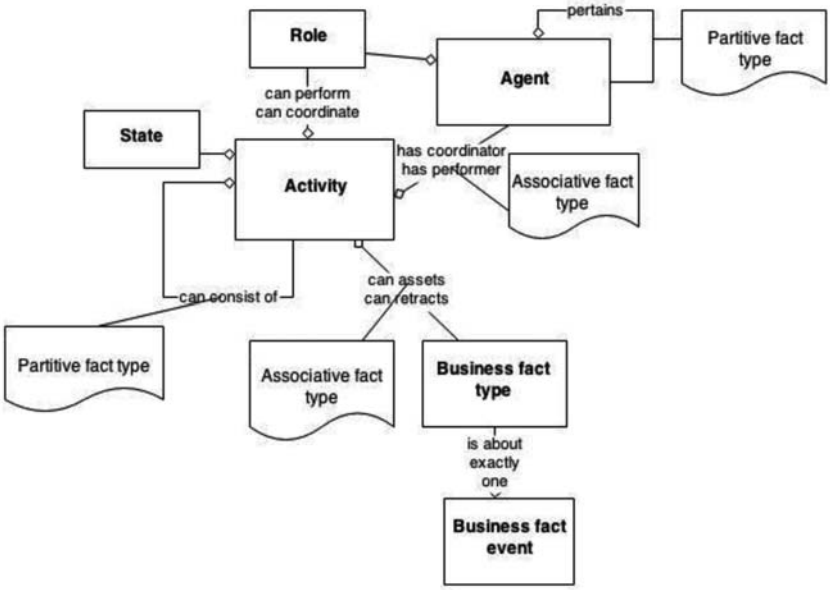

Fig. 1. Business concepts and fact types related to activity concept

Let's consider one essential concept from the EM-BrA2CE Vocabulary, for instance, the Activity concept. Fig. 2 displays the Activity concept representation in XMI, where the meaning of the Activity concept for Place order is underlined by a single line as it is a general concept type in SBVR.

```
<sbvr:conceptHasInstance concept="place_order_c"
   meaning = "meaning"/>
<sbvr:term xmi:id="place_order" signifier="activity_
order_s"
   meaning="place_order_c"/>
<sbvr:objectType xmi:id="place_order_c"/>
<sbvr:text xmi:id="place_order_s" value="Place
Order"/>
```

Fig. 2. Activity concept representation ion SBVR

The equivalent representation of this Activity concept can be expressed in OWL in Fig. 3, where the properties of the particular meaning of the concept are provided by the properties of an instance of a class `place_order`.

```
<owl:Class rdf:id="place_order"/>
```

Fig. 3. Activity concept representation ion OWL

This representation has a lot of advantages compared to their standard XMI expression in the SBVR standard. They can be summarized as follows:

- Readability – easy to read and understand.
- Support of data types – Support for handling values from common computing data types. Support of list structures.
- Object type hierarchies
- External data integration
- Built-in functions.

Each concept in SBVR has a respective instance that is employed in the execution of the business mode . For comparison let's consider the XMI representation of the instance of the Activity concept displayed in Fig. 2.

```
<sbvr:designation
xmi:id="Place_order_I signifier="Place_order_s"
meaning="Place_order_I_c"/>
<sbvr:individualConcept xmi:id="Place_order_I_c"/>
<sbvr:text xmi:id="Place_order_s" value="Place order
I"/>
<sbvr:concept1SpecializesConcept2
    concept1="Place_order_I_c" concept2="Place_order_
c"/>
```

Fig. 4. Activity instance representation in SBVR

Fig. 4 illustrates the XMI pattern used in SBVR to represent the *individual concept* Place order I , where double underlining in SBVR denotes the name for an individual concept of a concept type Place order.

The corresponding OWL representation of this individual concept is a RDFS instance type displayed in Fig. 5:

```
<rdf:description
   rdf:about="http://www.fmi.com/process#place_order_I">
  <rdf:type
   rdf:resource="http://www.fmi.com/process#pace_order"/>
</rdf:description>
```

Fig. 5. Individual concept representation in RDFS

Similarly, the individual concept for a Role can be represented in RDFS. Unlike a Place order individual concept, both Role individual concepts and State individual concepts have enumerated number of instances. Therefore the respective OWL representation of a Role individual concept

and a State individual concept takes the forms given, correspondingly in Fig. 6 and Fig. 7.

```
<owl:Class rdf:id = “Role”>
<owl:oneOf rdf:parseType=”Collection”>
   <owl:Thing rdf:about=”#Customer”/>
   <owl:Thing rdf:about=”#Publisher”/>
   <owl:Thing rdf:about=”#CreditManager”/>
</owl:oneOf>
</owl:Class>
```

Fig. 6. The Role individual concept representation in OWL

```
<owl:Class rdf:id = “State”>
  <owl:oneOf rdf:parseType=”Collection”>
    <owl:Thing rdf:about=”#Started ”/>
    <owl:Thing rdf:about=”#Assigned”/>
    <owl:Thing rdf:about=”#Completed”/>
  </owl:oneOf>
</owl:Class>
```

Fig. 7. State individual concept representation in OWL

This approach allows representing the remaining concept entities (Agent, Business fact types and States) displayed in Fig. 1 in terms of OWL and RDFS. Thus a complete hierarchical object oriented data model of the business Vocabulary can be obtained.

4 Business process model based on business rules.

The above proposed declarative representation of business vocabulary in terms of OWL and RDFS allows us to define a declarative hierarchical representation of a business process model. First we note, that the EM-BrA²CE Framework considers the business process model as a set of Activity types. The relationships among the elements of this set of Activity types is defined in terms of business rules of Event type. These business rules of Event type support a Start event business rule and appropriate Event types “fired” on Activity type initiation (precondition), execution and completion (postcondition). The EM-BrA²CE Framework additionally demonstrates that every Activity type can be modeled by describing its state space and the corresponding set of business rules that define the admissible state transitions in this state space. The set of the admissible state transitions in the state space of an Activity type actually represents the business process model for the execution of this Activity type. For clarity, we refer to the business rules defining the admissible state transitions in state space of a given

Activity as the Control flow of the business rules for this Activity. The Control flow of the Activity business rules of a business process can be presented in the form displayed Fig. 8.

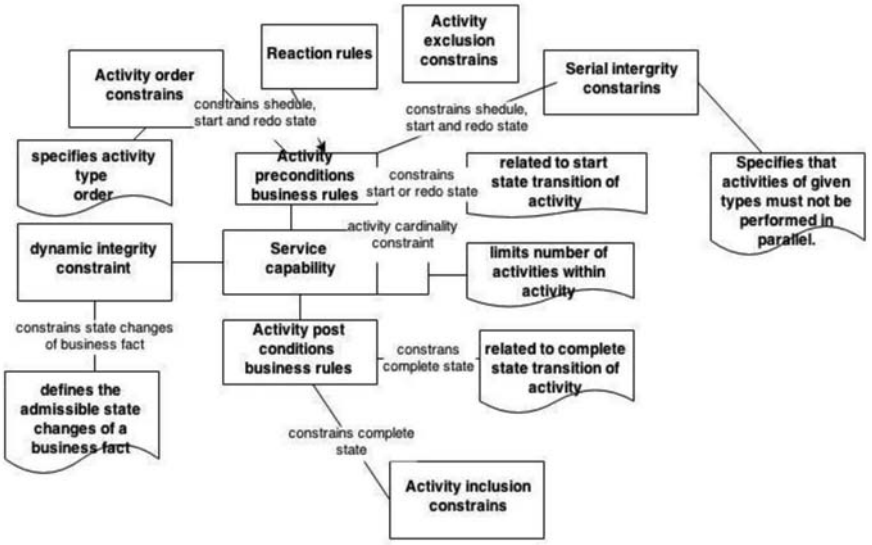

Fig. 8. Control flow of the Activity business rules

The Control flow of the business process involves reaction rule event processing, where actions are being invoked in response to events occurring in activity preconditions, post conditions or actionable situations in the activity execution. The recently adopted Reaction RuleML standard [14- 15] allows representing such rules in a hierarchical XML format compatible with the RDFS representations in Figs. 2- 7 of fundamental entities of the EM-BrA²CE Framework. Reaction rule events cause state changes during the lifecycle of an activity. A State change fact type instance in relationship to an activity is represented in Fig. 9 in terms of OWL and RDFS.

```
<owl:ObjectProperty rdf:id="StateChangeDateTime">
     <rdfs:domain rdf:resource="#Activity"/>
     <rdfs:range rdf:resource="#DateTime"/>
</owl:ObjectProperty>
```

Fig. 9. An Activity has State changed representation in OWL

The final step in obtaining a uniform hierarchical representation of a business process model is to relate the Rule set in the business process Control flow to the

Activity concept in the business vocabulary. Let's add an OWL Rule set class definition and OWL Activity class as superclass for every activity concept of the business process ontology as presented in Fig. 10.

```
<owl:Class rdf:id="Activity">
     <rdfs:label>Activity </rdfs:label>
     <rdfs:comment>Business activity definition
     </rdfs:comment>
</owl:Class>
<owl:Class rdf:id="RuleSet">
  <rdfs:subClassOf>
     <owl:Class rdf:id="Activity">
  </rdfs:subClassOf>
</owl:Class>
```

Fig. 10. A RuleSet and business Activity superclass representated in OWL

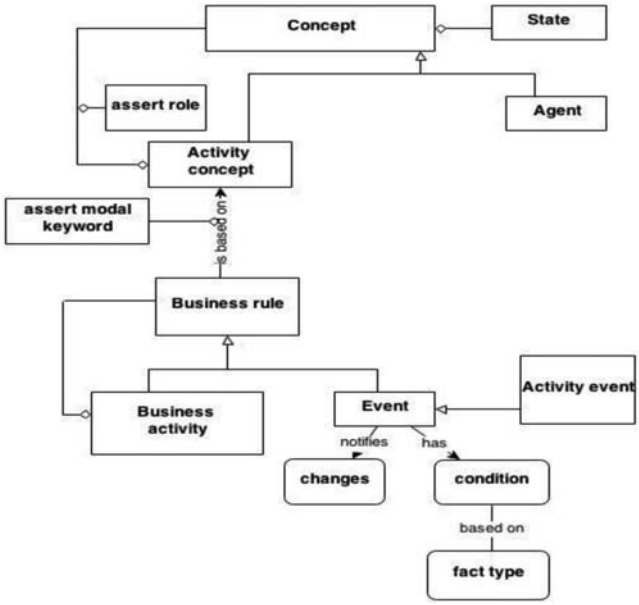

Fig. 11. Business process model based on business rules

An OWL Rule set class definition and relation to the Activity concept of the business process ontology as presented in Fig. 12.

```
<owl:ObjectProperty rdf:id="preconditions">
    <rdfs:domain rdf:resource="#Activity"/>
```

```
    <rdfs:range  rdf:resource="#RuleSet"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:id="postconditions">
    <rdfs:domain rdf:resource="#Activity">
    <rdfs:range rdf:resource="#RuleSet"/>
</owl:ObjectProperty>
```

Fig. 12. A RuleSet in relation to a business Activity representated in OWL

This allows us to refer to a Business activity as the RDF instance of the activity type represented by an OWL class. In the general case the Activity Business Concept has to be extended by means of the Composite design pattern, where a business rule represents a reaction rule or a business activity. On the other side, the Composite design pattern represents the business activity as an aggregation of business rules. The declarative representations of the basic rules in the EM-BrA²CE Framework for Activity related concepts and their instances allow us to propose a uniform hierarchical model for declarative representation of a business process displayed on Fig. 11. The thus obtained declarative representation of a business process integrates the business vocabulary and the business rules controlling the flow of events, respectively, the state changes in the execution of activities.

5 Conclusion.

The declarative representation of a business process model reflects the convergence in emerging standards for modeling business processes and business rules. This paper presents an object-oriented approach for transforming the EM-BrA²CE Framework set of rules into a declarative rule- based representation of a business process model employing OWL and RDFS standards. The recently OMG adopted Rule ML interchange format allows the thus obtained business process model to be tested and verified by standard rule execution engines.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Program under contracts BG051PO001-4.3.04-0018(2013/2014), BG051PO001-3.1.08-0010 (2012/2013).

References

1. Jakob Freund, Bernd Rücker, “Real-Life BPMN. Using BPMN 2.0 to Analyze, Improve, and Automate Processes in Your Company”. Camunda, 2012.
2. Object Management Group. “Business Process Model and Notation (BPMN) 2.0.2” http://www.omg.org/spec/BPMN/2.0.2/PDF/, August 2014(accessed Aug 10, 2014)
3. Jan Mendling, Markus Nüttgens, “EPC markup language (EPML): an XML-based Interchange format for event-driven process chains (EPC)”. Information Systems and e-Business Management, Volume 4, Issue 3, pp. 245-26, .July 2006.
4. Jan Recker et al., “Do Process Modelling Techniques Get Better? A Comparative Ontological Analysis of BPMN”. Procs. of the 16th Australasian Conference on Information Systems, November 30- December 2, Sydney, Australia (2005)
5. Vid Prezel et.al.,”Representational Analysis of Business Process and Business Rule Languages” , Procs. of 1st International Workshop BuRO (2010)
6. Michael Rosemann et al., “A Study of the Evolution of the Representational Capabilities of Process Modeling Grammars”, Procs. of CAiSE, Springer, Luxembourg, pp. 447-461. (2006)
7. M. Muehlen et. al., “Business Process and Business Rule Modeling Languages for Compliance Management: A representational analysis”, Procs. of the Twenty-Sixth International Conference on Conceptual Modeling , New Zealand 2007
8. M. Muehlen Marta Indulska, “Modeling languages for business processes and business rules: A representational analysis”, Information Systems vol. 35 (Issue 4, 2010)
9. Ronald G. Ross,”Business Rule Concepts: Getting to the Point of Knowledge” (4th Ed.), Business Rule Solutions, 2013
10. Ross, R. (ed.) and the Business Rules Group., “Business Rules Manifesto: The Principles of Rule Independence.”, 2003, http://www.businessrulesgroup.org/brmanifesto/BRManifesto.pdf
11. Semantics of Business Vocabulary and Business Rules (SBVR) Version 1.2 Specification, http://www.omg.org/spec/SBVR/1.2/, OMG, 2013
12. S.Goedertier et al., “Rule-based business process modeling and enactment”, Int. J. Business Process Integration and Management, Vol. 3, No. 3, pp.194–207 (2008)
13. Jaroslav Karpovič, Gintarė Kriščiūnienė, Linas Ablonskis, Lina Nemuraitė ”The Comprehensive Mapping of Semantics of Business Vocabulary and Business Rules (SBVR) to OWL 2 Ontologies,” 2014
14. Adrian Paschke et.al., “Standards for Complex Event Processing and Reaction Rules”, RuleML 2011 - America, LNCS 7018, pp. 128-139, Springer-Verlag Berlin Heidelberg 2011
15. Adrian Paschke et.al., “Specification Reaction RuleML 1.0”, http://wiki.ruleml.org/index.php/Reaction_RuleML (2013)

The Design and Realization of the Municipal Informational and Administrative Website

Krasimir Nikolov[1] and Svetlana Vasileva[1]

[1] Shumen University „Bishop Konstantin Preslavski“, College – Dobrich,
Dobrotitsa 12, Dobrich, 9302, Bulgaria
{1sun_7, 2svetlanaeli }@abv.bg

Abstract: The Content management systems (CMS) automates and facilitates the process of adding and modifying the contents of the Web sites, organization, control and publication of a large number of documents and other content, such as images and multimedia resources. This makes the Content Management Systems attractive for specialists in various fields of human activity, who want to publish on the Internet, but have little knowledge in computer programming and in particular web-programming. This paper discusses the problems in the creation and implementation of web-based information systems for municipal administrative centers. The information issued by the information system is the link between the citizens, businesses and municipal government. The paper presents a web-based information system serving the “Center for Business and Culture - AD” - Dobrich.

Keywords: municipal administrative activities, content management system, information system.

1 Introduction

The national targets to achieve efficiency in the information systems are related with the development of e-government in Bulgaria. Such a structure is required in order to include the country in the European information infrastructure.

By adopting the e-Government Strategy, Bulgarian Government commits itself with the development of 20 indicative services performed electronically - 12 for citizens and 8 for businesses. [5]

The plan to implement the strategy provides specific projects related to the implementation of the indicative services. Administrations that are responsible for the performance of each service, have been identified. The administrations are obliged to take concrete action to promote their services on the Internet and encourage the use of services over the Internet.

The introduction of the European Commission’s 20 indicative electronic services for businesses and citizens provide four levels of service delivery [5]:

First stage - Information: institutions publish the information online that is accessible to citizens and businesses;

Second stage - One Way interaction: institutions publish the information on the Internet and enable the downloading of blanks and forms relating to services;

Third stage - Two-way interaction: the user of the service besides receiving information and downloading blanks can send letters, forms and more, electronically, but the administration is not obliged to answer in real time or in the same way;

Fourth stage - Transaction: citizens and businesses communicate with the administration electronically and vice versa "online". There is a mechanism to verify the validity of the transaction.

These 20 recommended services do not restrict agencies to introduce additional services. The central authorities have developed and offer consumers the following electronic administrative services other than 20 indicative ones [5]: The Registry Agency; Registration of businesses in BULSTAT register; Property registration; Sending data to the central authorities (MVR, NRA, NCA, etc.); Department of Defence; Public procurements; Complaints Accreditation of the journalists; Department of Economy and Energy; Public procurements; Department of Education and Science; Register of secondary schools and kindergartens; Register of higher education institutions; Reference system for municipalities; Geographical reference system of the structure of public education; Ministry of Justice; Conviction status certificates; Department of Labor and Social Policy; Requests, complaints, signals; Access to public information; MLSP structures; Social Assistance Agency; Administrative services - requests, complaints, signals; Access to public information; Social services; Family allowances for children; The General Labour Inspectorate; Declarations, requests, complaints, warnings; Permissions in chl.302, chl.303 and Article 333 of the Labour Code; State Agency for Child Protection; License social services; Employment Agency; Search for job openings.

Electronic services provided by public institutions could clearly be defined as business services. [5] Avoiding direct contact with the administration, where possible, and replacing it with a virtual one, is an attractive alternative for companies. Network becomes increasingly preferred in obtaining and providing information and documents when making financial transactions. Third of managers in right of option, would communicate with the institutions without leaving their workplace. Unlike businesses, to the population is more difficult to give up face-to-face communication and many more prefer the traditional working the counter. There is a distrust of Internet communication, especially when it comes to paperwork and financial transactions. One possible reason for this is that "web-environment is relatively" young "and still not well utilized by both the creators and consumers of web-sites" [9]. And this in today's globalized world is an obstacle to business development overall in Bulgaria. The whole work on web

based information system of “Center for Business and Culture” SA is the initial part of what is needed to build a full face of this organization. From creating the database to the site design, layout and menus shaping the documentary part of the development can help create a large project with practical action.

To create the web based information system is no need to use certain software. Most suitable financial terms are the content management systems (CMS) a free license, providing powerful enough funds to build a dynamic site with a link to the database.

2 E-government and E-municipality

Fundamental principle when providing information and services is “equal access for all.” This means that all channels of access to information and services must be effective, consistent and meet the same standards - unified legislative basis, independent of the various technology services. Internet can not and should not replace other ways and means of access to the services of the administration.

E-government is (despite the overall technical and telecommunications environment) is set of relatively autonomous systems that perform three main and one auxiliary function [5]:

A. Function “macro-management of the state” in which the predominant analytical and synthetic procedures related to the processing of unstructured information, in advance unformulated outputs and long time periods to generate solutions.

B. Function “E-services for citizens and businesses,” in which predominate formalized procedures for handling structured information in a mode close to real time.

C. Function “Exchange of information between departments of the administration” - the exchange of data, primarily related to the technological processes within the administration. This includes the exchange of structured data; exchange of unstructured data (including graphic-organized data and multimedia); exchange of metadata.

Primary means of achieving the objectives of e-government are the new information technologies, developed on the basis of the Internet.

In the information society E-government is closely connected with the development of the E-municipalities. The municipalities in Bulgaria have separate administrative role, which is not derived from the central government. [9] “In the draft E-government the municipalities should not necessarily expect urge the central government - municipalities have their own role must fulfill” [10]. The review of the state of the municipal websites in Bulgaria shows that the web-sites of local government are very different from one another, is in some of it difficulty to find the necessary information or access to electronic service.

There are differences in the number and manner of providing e-services. All this raises the problem of making recommendations not only (as specified in [10]), but also research and preparing projects and strategies similar to those of E-government. In other words, the E-municipality should become a local version of the E-government.

3 Structure of web-site

The website is an Internet presence, a collection of texts and images, sound and animation together in pages available to visitors through an address in browser. [7] Good organization of pages unites them on one hand meaningful by the same style of the texts, references and inspiration, on the other hand by a single graphic colors, graphic elements, similar processing of pictures, effects, and more. The use of a single style in his address to the consumers of text level, and the use of graphical detail similar or identical filters in the processing of pictures level vision creates a sense of integrity - making the site popular, stylish and pleasant to use. The site has hierarchical levels homepage, entry page for each module and internal pages. [7], [11] All connections between pages of different levels, and the overall functionality of the site should be considered in advance and to set the design. [7] [12] It is important to clarify its structure and architecture and to establish the format of each level. In some cases, selecting different colors for various sublevels. Intuitively understandable for most users is a pyramid structure in which there is a central entrance leading to several sublevels referring users more “deeper” site. Is created a logical and easy to understand hierarchical structure, in which the visitor moves easily from external levels containing more general information towards internal levels where the specific data are and vice versa. However, to avoid unnecessary traffic to the site, are created direct links between the levels and binding on all pages a link to the so-called additional pages. Between pages on the site there are many links which are not always subject to the hierarchical principle.

The main elements of each site are: Home; Major internal levels; Internal pages. [7]

Additional elements: Site Map; Search form; Contact form; Frequently Asked Questions (FAQ).

Planning the site structure standard [6], [8] ends with the preparation of a flowchart with different levels, additional pages and links between them.

3.1 Characteristics of the construction of Municipal Informational and Administrative Websites

Main steps for the implementation of Informational and administrative site are: Stage I - online information about public services. Stage II - opportunities for

citizens and businesses to use and fill out electronic forms for the requested services. Stage III - Processing completed electronic blanks requests for public services, authentication (including electronic signature). Stage IV - processing and resolving requests for public service delivery and payment of the requested service.

Fig. 1 shows the Home page of the Dobrich municipality informational and administrative web site [3]. Fig. 2 shows the Page of the Center for services and information of the Dobrich municipality.

Fig. 1. Home page of the Dobrich municipality informational and administrative web site.

Fig. 2. Page of the center for services and information of the Dobrich municipality.

According to [9] the website of the Dobrich municipality is good practice (sparingly menu, the existence of the “What’s New”), but we think the Dobrich

municipality site, does not facilitate the users if they have to turn to one of the sub-menus. Moreover, somehow at the center of the site is loaded the realization of the „infotainment" functions of the site. Users with inadequate information culture, but who want to seek information and to receive e-service on the website of the municipality will find themselves difficult or will lose time to find what they need. Often the user hurries and the indirect access additional makes him nervous. From this perspective, the site of the Burgas municipality [2] (fig. 3 and fig.4) is much more open, and the e-services (fig. 5) offered by Burgas E-municipality are more than Dobrich.

Fig. 3. Home page of the Burgas municipality informational and administrative web site (I).

Fig. 4. Home page of the Burgas municipality informational and administrative web site (II).

The ordinary citizen (in our view of the Burgas site) will easily find the

necessary information or the e-service menu on the left of the screen (enough for the user to scroll down the screen - fig. 3 and fig. 4).

Fig. 5. E-services page of the burgas municipality informational and administrative web site.

3.2 Design of an Informational and Administrative website

The information system of an information and administrative site is essentially a database of services and opportunities for doing business in an administrative unit such as a municipality, it may seek different types of information from which the given user is interested.

The database of the projected site contains four tables. The table “Contacts” contains fields describing the address and any way to connect with “CBK” AD. The table “Projects” contains information about projects “CBK” AD yet to realize. The table “About Us” contains information about the structure and composition of the governing body of the “CBK” AD, as well as information about the company since its inception, respectively table “Services” describes all the services and capabilities of the “CBK” AD. CMS Drupal “read” the code and creates the tables and the database (to its database). In variables schema describes the fields of the tables. CMS Drupal “know” that set forth in this variable are described in tables and use system functions to implement the tables, and then described relations between tables.

4 Features of Content Management System Drupal

Web-based content management systems are used for the preservation and publication of documents. [1] The open systems are established, maintained and developed by many developers. Their code is publicly available for reading and editing. This provides greater flexibility, stability, and a variety of additional

modules and possibility for their functionality extend. CMS allow the creators to be independent of web design companies and are able to update and modify the content of the web sites. [1], [4], and [9] Each CMS could be appropriate in some conditions and inappropriate for others. Choosing a CMS should be dictated by the nature and needs of the site for which it was intended.

The Content management system Drupal is a mature system with enormous opportunities. It is a free, powerful and popular, and is also open. Drupal architecture allows for a complete various types of web sites, including educational sites. Existing functionality by default can be increased by connecting different extensions - "modules" in the terminology of Drupal. These additions provide a full range of features that make the system very robust and easy to use CMS.

As a content management system Drupal provides [1], [9] the following features of the:

Creation of documents and multimedia materials;

Identification of all key users and their roles in the content management;

Ability to assign roles and rights of the different users of different types or categories of content;

Manage workflow to create content: it is a process of creating cycles of sequential and parallel tasks, which have to be fulfilled in the content management system. For example, the author of the article content added, but it is not published until the editor does not check and the editor did not approve it;

Ability to publish content in the mining and access to it;

Automated templates: created by the system and can be automatically applied to new or existing content and their change affects the appearance of all pages of the site;

Content which was edited: immediately after the separation of the content of the visual representation of the site it is generally more susceptible to manipulation and editing. Most CMS include WYSIWYG tools for editing, allowing non-technical staff to create and edit content;

Simplified adding new capabilities: Most CMS have plug-ins or modules that can be installed easily and can extend the existing functionality of the site;

Constant updates. Most CMS usually offer such upgrades incorporating new features and support system with the latest web standards.

CMS Drupal has enough power and flexibility, allowing us to create a topic that is complex enough. The system offers countless ways to deal with problems that arise, but you need to know how to work with Drupal themes so to choose the proper way.

5 Realization of Informational and Administrative Websites by CMS Drupal

Fig. 6 shows the homepage of the informational and administrative web site of „CBK“ AD at the stage of design in the environment of Drupal. On the left side of the window are the fields for quick search and advanced search.

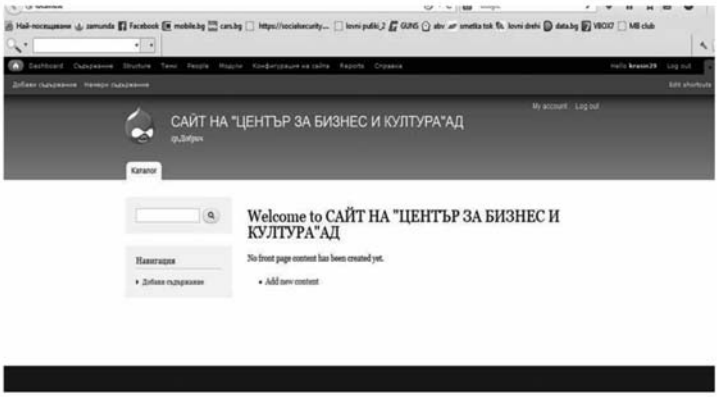

Fig. 6. E-services page of the CBC”AD informational and administrative web site.

FIG. 7 shows an example query execution for fast keyword search “contracts”. Fig. 8 shows a window with results from the application of the example in Fig. 7.

Fig. 7. Query execution for fast keyword search “contracts” in the informational and administrative site of „CBK”AD.

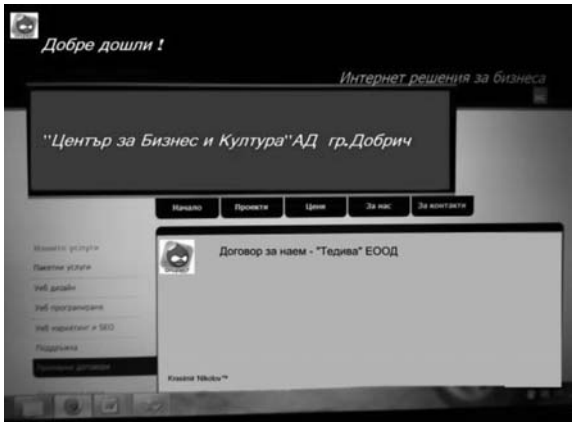

Fig. 8. Page with the results from a quick search in the website of "CBK"AD.

6 Conclusion

The information about the status of municipal informational and administrative web sites in Bulgaria is not enough. In addition there are very few the municipal informational and administrative web sites, fully satisfying the citizens of the municipalities and companies.

The made review of the municipal websites showed that there is no standard for the implementation of informational and administrative web sites of municipalities in Bulgaria. That "in today's" global world is an obstacle to the development of the E-government in Bulgaria.

CMS Drupal is suitable for the functional and professional platform for Municipal Informational Administrative Web site. (Do not forget that many government sites are realized by means of CMS Drupal!). With the development and implementation of Informational Administrative Web site of the Center for business and Culture - AD, Dobrich hope to grow into a project to improve municipal e-services as part of the development of the E-municipality of Dobrich.

References

1. A review of open source content management systems. - http://www.openadvantage.org/articles/oadocument.2005-04-19.0329097790 (2005)
2. Burgas municipality: Official website. http://www.burgas.bg/
3. Dobrich municipality: Official website. http://www.dobrich.bg/
4. Drupal API. https://api.drupal.org/api/drupal
5. E-government in Bulgaria, izt.bgdocs.org/docs/index-234972.html (2013)
6. Information page on the Quality, http://www.ncbi.nlm.nih.gov.
7. Kotcev, M.: Web design and web site. Web-site structure, http://content-management-systems.info/node/619
8. National Centre of Information and Documentation. http://mail.nacid.bg/newdesign/bg/index.php?id=68
9. Reference book in Drupal. http://content-management-systems.info/node/619
10. Sirkov, I. Usability on municipal websites in Bulgaria - achievements and upcoming tasks. Foundation „Applied Researches and Communications". www.lucrat.net/content/1_ARC_Fund_Obshtini.ppt
11. Web-site planning. Internet solutions for the business. http://designforce.bg/bg/articles-internet-marketing/36-manageyeb-2
12. Web design, http://www.escom.bg/?module=bussines&action=details&id=18

* This paper is supported by the Project BG051PO00l-3.3.06-0003 "Building and steady development of PhD students, post-PhD and young scientists in the areas of the natural, technical and mathematical sciences". The Project is realized by the financial support of the Operative Program "Development of the human resources" of the European social found of the European Union.

Analysis of Business Process Models

Kristiyan Shahinyan [1], Evgeniy Krastev [2]

[1] ComSoft Ltd., 47, Knyaginya Maria-Luiza blvd., floor 1, 1202 Sofia, Bulgaria,
[2] Faculty of Mathematics and Informatics, St. Kl. Ohridski University of Sofia,
5 James Bourchier Blvd., 1164 Sofia, Bulgaria
1 k.shahinian@comsoft.bg, 2 eck@fmi.uni-sofia.bg

Abstract. Business process modeling is an essential part of both organizational design and information systems development. It allows ignoring irrelevant complexities of a world driven by a continuous change of requirements and advances in technologies, while keeping the focus on the most important parts of the system under study. The inherent complexity of business process models brings up the problems of analyzing their syntax and logical consistency, subject to the selected modeling technique. The execution of a business process model raises the problem of evaluation of certain metrics in accordance with Key Performance Indicators (KPIs). This paper presents an approach to analyze both the static and the dynamic aspect of a business process model. Unlike other similar approaches this approach takes into consideration all the business processes comprising the activities in a business organization subject to modeling. A realistic case study demonstrates the feasibility of the proposed software implementation of this approach in both static and dynamic analysis of complex business models.

Keywords: Business process model, static analysis, dynamic analysis, EPC, BPMN.

1 Introduction

Business process analysis is an important discipline in modern organizations. Its main goal is to identify business needs and provide solutions to business problems. Most commonly business analysis results in description of analyzed processed. These descriptions are provided in graphical notations such as Flowcharts, Integration Definition (IDEF), Petri Nets [1]

Flowcharts are one of the first graphical modeling techniques. They document the overall structure of a system and shows the information flows. Examples of flowcharts are an Event-Driven process Chain (EPC) and BPMN. IDEF can be divided into three independent techniques, function modeling, data modeling and process description modeling. The techniques model processes and data structures in an integrated fashion. Petri nets provide a more mathematical/graphical model of systems to analyze the structure and dynamic behavior of modeled systems.

Every standard/notation has its area of relevance. They are used for different purposes and aim different goals.

Any of the fore mentioned notations are used nowadays, but EPC and BPMN are imposed in practice. Therefore, in this paper we will consider the BPMN 2.0 [2] and EPC [3] representations of a business process model. We will perform static and dynamic analysis on these definitions and provide interpretation of the results.

2 Problem statement

When taking into account a single process, its graphical notation is sufficient in order to define required resources, generated and used documents, related information systems and so on. On the other hand analyzing a single process is not enough when analyzing company's strategy Key Performance Indicators (KPIs). In this case we should consider all process definitions. This also includes changes in processes and monitoring of business process instances.

When changing a business process we should take into account its semantic correctness (according to its business notation), its correctness according to any other business process in the organization and its correctness according to process execution. Semantic correctness should be evaluated having in mind not only the business notation but also some logical requirements [3], [4]. Process correctness might be reconsidered when there are some organization requirements such as custom notation extensions. Process definition should also be examined with accordance to its execution. This is even more important when process is supposed to be (partly) automated. This includes analysis in respect of execution engine and its requirements [5], [6].

When analyzing the process as a definition we perform a static analysis. Static analysis is expected to claim the process correctness in accordance to notation's and organization's requirements. In static analysis we analyze all resources and artifacts related to the process, their availability in the organization's infrastructure and the requirements for all information systems, used by the process. When performing static analyses we should not analyze only the process itself, but we should analyze the whole set of business processes, because those processes most likely share common resources (like employees, information systems, documents, etc.).

On the other hand we should also analyze the process execution, using execution monitoring, execution simulation, etc. We call this dynamic analysis. This analysis could give us hints what we could change in order to achieve some organization's goals. For example we can perform what-if analysis or integrate a brand new process in the execution environment controlling the level of interception with existing ones. We should perform dynamic analysis when

we are trying to create an execution environment for those processes. Otherwise this analysis will result in false positive or absolutely incorrect results. Common practice in dynamic analysis is to use empirical results (extracted from day-to-day execution) for every activity in the process.

Both types of business process analysis should be executed over a set of business processes. Analyzing a single process is not enough, because we could miss some resources collisions (like employees or documents). On the other hand we should provide some requirements for information systems, invoked by the process. If this information system is used by another process we could create different interfaces and therefore issues in real-time execution.

3 Static model analysis

Static model analysis provides a good overview of business requirements, business needs and process complexity. Static model analysis is quite useful when analyzing a set of business processes, having common resources (for example, one process generates a document and a few others use the same document). This analysis approach provides better understanding for all artifacts and resources, used in an organization.

In this paper, we describe an automated approach for static model analysis of EPC diagrams, where the same approach is applicable to static model analysis of BPMN diagrams. For every diagram, a custom developed software tool extracts all the functions, events and artifacts (generated documents, used documents, associated information systems and user roles). After the extraction is complete, the software performs a few cross-reference checks for the artifacts in all models. For example, it matches generated documents from every diagram with the documents used in all the diagrams of the business process model subject to investigation. When there is, a document is generated but it is not used anywhere, a potential issue is identified and it has to be investigated thoroughly. In general, it is pointless to generate documents that are of no use. Nevertheless, there might be "False-Positive" results, for example, such a document might be generated for an external organization. The software also checks a few "best-practice" templates, for example, a function should not be followed by another function (instead an event or a condition should be present) [3]).

When performing static analysis we should keep in mind the following evaluation metrics:

- Complexity– business processes should be readable for "business users". They should be an image of the steps, followed for achieving a certain "value-added" result. That is why business processes should be kept simple as much as possible, easy to follow (execute) and understandable;

- Correctness– business process models always follow a notation (or a standard). All notations define components and relations and rules for their composition. This composition (as a result this is a business process model) should be syntactic (following notation requirements and definitions) and logical (following business organization's logic) correct;
- Model scope– the scope of every business process should be defined in advance, based on the scope of a set of business processes. This set is supposed to perform an organization's daily tasks using a composition of the processes included in the set.

Thus the static analysis provides a couple of useful reports:

- Appliance of models with best practices, taken into account by the organization
- Generation of documents and their usage in the organization;
- A complete list of the responsibilities of the staff involved in the business process model execution
- Departments' communication, the exchange of documents used in different departments as part of their interaction during the business process model execution
- Information systems usage, describes which information systems are used by a department (and even employee), what do they contain, how they integrate with other information systems and other similar characteristics.

4 Dynamic model analysis

The Dynamic model analysis includes all the activities performed in a "production" business process execution. It is not necessary to automate all such activities or provide the information systems necessary for the process execution. For example we can monitor and analyze on a sheet of paper the execution of a particular business process, the payment of a term fee at a university. Monitoring the process this way could be quite a complex task. It involves gathering all the information, systematization of all data and so on. Everything could be different if we use a process execution environment. Such a software application could extract all the information we need and provide additional analysis operations, for example, providing a standard SQL interface for data analysis and even Business Improvement analysis. In order to achieve this we should be able to define some key process activities or conditions that should be analyzed (for example, monitor activity execution time or frequency). A custom software tool based on the Activiti Framework [5] is being developed and used in this paper for automated simulation of a BPMN 2.0 business process.

It is useful to employ a simulation environment in order to provide "production" ready business process models. In this environment, we can define and evaluate metrics for evaluating the execution time of activities, processes or related processes, detect bottlenecks in a process or process sequence execution. Most importantly, we can define metrics and quantify modeled activities, sub-processes, processes or a set of related processes. Accordingly, we can use the following relation:

$$\text{Cost} = F(\text{Time}, \text{Quality}) \quad (1)$$

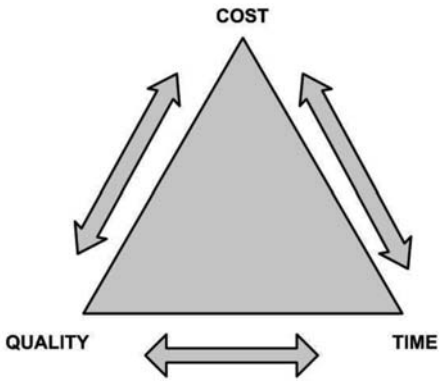

Fig. 1. The Time- Cost- Quality relaiton

The Time- Cost- Quality relation (1) suggests that the business process model evaluation is a subject of trade-offs of metrics related to these three characteristics. Depending on the Key Performance Indicators (KPI), two of these characteristics are selected as independent subjects. Evaluation metrics belonging to the independent subjects determine and optimize evaluation metrics belonging to the remaining subject in the relation (1) by assigning desired bounds or particular values on the independent subjects. Such an approach to dynamic model analysis allows us to describe the quality of the business process model by identifying its displacement in **Fig. 1.** The Time- Cost- Quality relation. Once the KPI define the target position of the business process model on Fig. 1, simulation results obtained from related evaluation metrics measurements serve to adjust the business process model execution to its target position characteristics.

5 Experimental results

Let's consider a common business process at a university, the process of generating a Summary report for research activities at the university. For clarity, we have documented the sequence of activities and the whole business process as it is performed at Sofia University. This report is generated on a yearly basis

and summarizes all the research achievements, publications, completed projects, presentations on scientific conferences and so on, of all the lecturers and researchers at the university. Some of the faculties use an Information system, named "The Authors", which stores this kind of data and reuses it in generating reports similar to the Summary report for research activities at the university. Other faculties collect and process manually this data to prepare such a report. A high- level BPMN diagram of the business process is depicted on **Fig. 2**. The Summary report for research activities is compiled by the Department for Scientific and Applied Activities in the headquarters of the university. The business process starts by a request issued from this department to all the faculties at the university. Next, each one of the faculties summarizes the data received from the lecturers employed in its departments. Finally, the summary reports from all the faculties are being forwarded to the Department for Scientific and Applied Activities in the headquarters, where the Summary report for research activities at the university is being worked out. Note, that each one of the faculties prepares its Summary report for research activities in parallel with the rest of the faculties. The same refers to all the departments at the university.

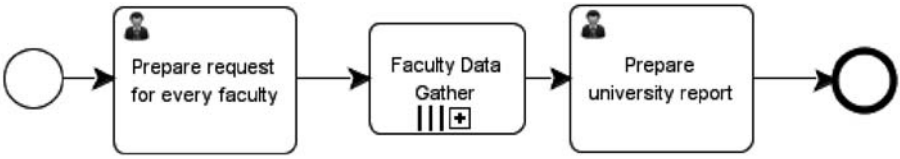

Fig. 2. High- level BPMN diagram

The activity, named "Faculty Data Gather" in **Fig. 2**, is modeled as a subprocess (*defined by the plus mark in the middle*) with parallel executing multi-instances (*defined by the three vertical lines*) of that subprocess, where a detailed process definition is shown on **Fig. 3**). When a request to prepare the Summary report is received at a faculty, then there are two options. If the information system "The Authors" is available, then the report is automatically generated by the information system. Otherwise a request to every department at the respective faculty is being sent. When all departments submit their report, then all the reports get merged into the Summary faculty report.

The "Faculty Data Gather" subprocess expansion is shown on **Fig. 3**, where the activity, named "Department Data Gather", is being introduced. This activity is modeled as a multi-instanced subprocess, executed by every department in every faculty. When a department receives a request, a request is sent to each one of the lecturers. All the lecturer responses are summarized in the department report. The subprocess "Prepare Personal Report" is depicted on **Fig. 3** and executed as a multi-instance subprocess, as well.

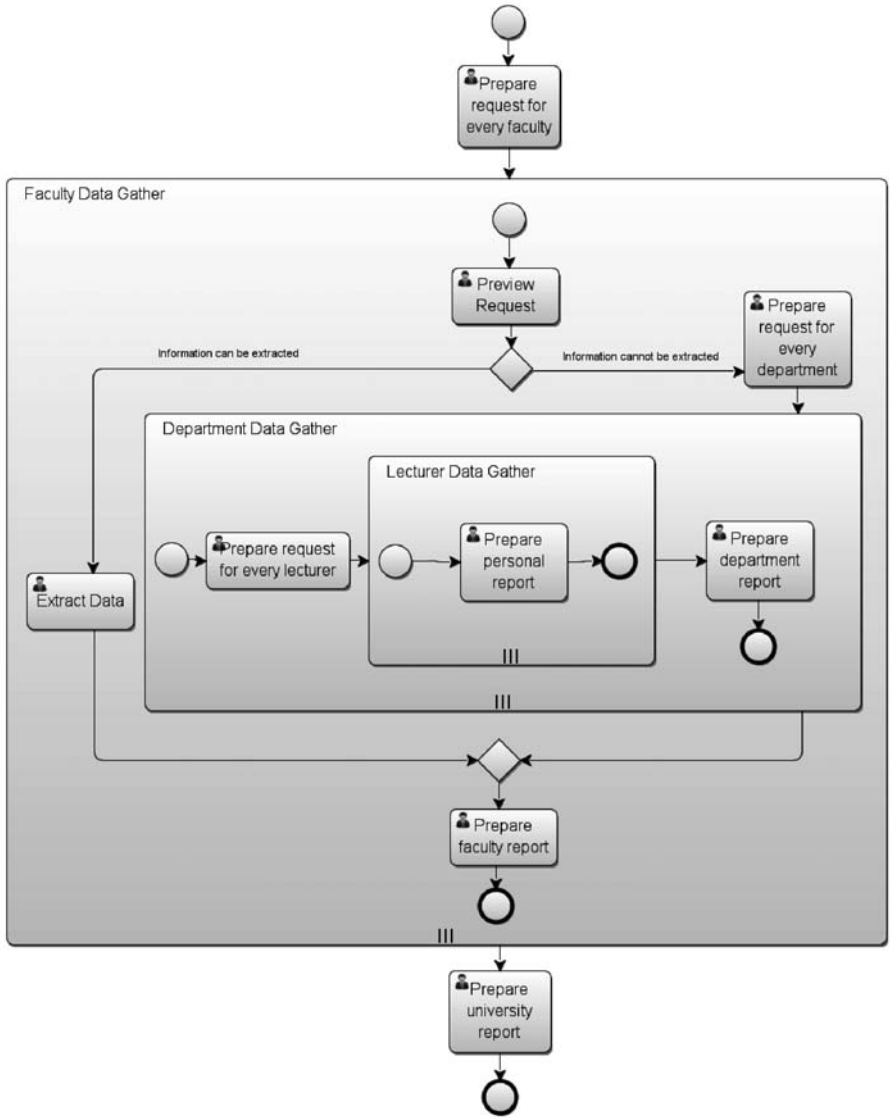

Fig. 3. Detailed process definition

The here considered business process is shown as a BPMN 2.0 on **Fig. 2** and **Fig. 3**. Let's consider now the same process definition in AML (Aris Markup Language) on **Fig. 4**- **Fig. 6**. The custom developed software tool, using the here proposed approach for static analyses of an AML modeled business process, generates a process report. The static analysis report of the business process "Generating a Summary report for research activities at the university" is shown on **Fig. 7**. Note

that the "Annual scientific report, containing results, containing results, scientific cooperation, resources and staff" is generated in **Fig. 6**, but it is not used in the business process. It is an example for a "False positive" result, because it is generated with the purpose to be used by an external organization. Similar "False positive" result is the Application form of the Ministry of Education for reporting of scientific activities, that is used in **Fig. 4**, while it is not generated as part of the business process. This document is provided by an external organization. On the other side the existence of any other documents that appear not to be "False positive" results are an indication for an incompleteness or logical inconsistency in the business process model. Accordingly, the business process model has to be revised in such case.

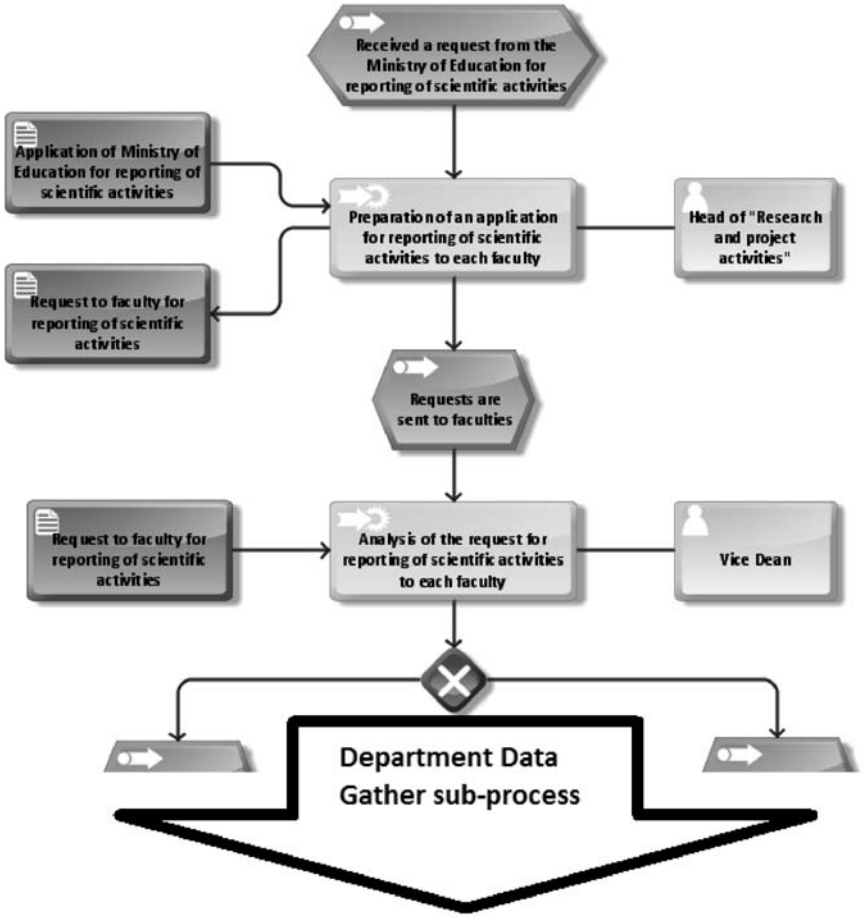

Fig. 4. Process "Generating a Summary report for research activities at the university"– Part 1

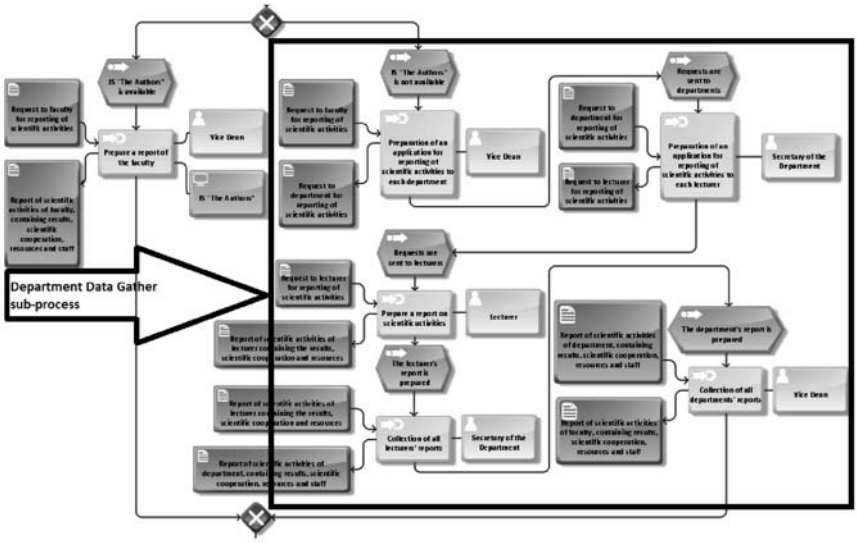

Fig. 5. Process "Generating a Summary report for research activities at the university" – Part 2

Now, let's consider a dynamic analysis of the same business process by taking into account the KPIs for operational performance [4], namely, the productivity and workforce turnover with respect to the above described BPMN model. These indicators depend on the "cycle time" performance indicator. It is common to use evaluation metrics in such cases [5]. Similarly, in order to measure this indicator we introduce the following evaluation metrics:

- The Total count of an activity execution defines how many times a single activity is executed in a single process execution instance
- The Total time for executing an activity and a subprocess – this defines how much time is needed for parallel execution of every activity and subprocess in a single instance of the process

The Average time for executing an activity and a subprocess – this defines the average time needed for parallel execution of every activity and subprocess

In computing [6] the time related evaluation metrics we refer to the following benchmarks for manual execution of activities in report generation:

- The Minimum time measured in minutes, required for completing an activity
- The Maximum time in minutes, required for completing an activity.

The benchmarks we have used have been coordinated with employees involved in the real-life execution of the process.

Table 1 and **Table 2** show the obtained values of the evaluation metrics for "cycle time" (time metrics are in minutes), obtained from the simulations of the

BPMN process model in **Fig. 3**. The metrics measured are the total execution count, the total time and the average time.

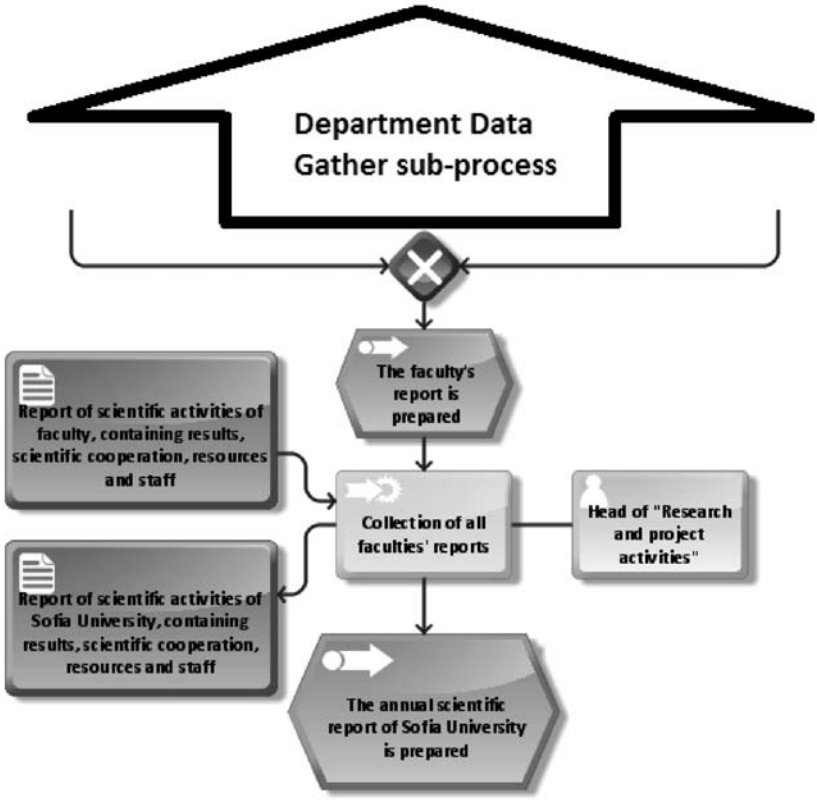

Fig. 6. Process "Generating a Summary report for research activities at the university" – Part 3

For every evaluation metric there are four dimensions, where the abbreviations used in **Table 1** and **Table 2** are shown in brackets:

- Best case execution (B) – in this simulation we are taking the beginning of defined time intervals, i.e. the shortest time for an activity execution;
- Worst case execution (W) – in this simulation we are taking the end of defined time intervals, i.e. the longest time for an activity execution;
- Random execution (R) – in this simulation we are taking a random numbers from each activity time interval, i.e. the activity is executed between the minimum and the maximum of this interval.;
- Perfect execution (P) – suppose all faculties are able to prepare their reports from external information system.

Generated documents:
Request to department for reporting of scientific activities
Request to lecturers for reporting of scientific activities
Request to faculty for reporting of scientific activities
Annual scientific report, containing results, containing results, scientific cooperation, resources and staff
Report of scientific activities of department, containing results, scientific cooperation, resources and staff
Report of scientific activities of faculty, containing results, scientific cooperation, resources and staff
Report of scientific activities of lecturer containing the results, scientific cooperation and resources
Generated documents are not used:
Annual scientific report, containing results, containing results, scientific cooperation, resources and staff
Used documents:
Request to department for reporting of scientific activities
Request to faculty for reporting of scientific activities
Request to lecturers for reporting of scientific activities
Application form of Ministry of Education for reporting of scientific activities
Report of scientific activities of department, containing results, scientific cooperation, resources and staff
Report of scientific activities of faculty, containing results, scientific cooperation, resources and staff
Report of scientific activities of lecturers containing the results, scientific cooperation and resources
Used documents that are not generated:
Application form of the Ministry of Education for reporting of scientific activities
Information Systems:
IS "Authors"
Employees:
Vice Dean
Head of "Research and project activities"
Lecturer
Secretary of the Department

Fig. 7. Static analysis for the Preparation of annual scientific report

The evaluation metrics Total time for executing an activity and Average time for executing an activity in the case of the above four dimensions are shown in **Table 1** and **Table 2**. All of the simulations have the same input variables:

- Number of faculties at the university– 16 (fixed);
- Number of faculties using "The Authors" information system for extracting report data– only 1;
- Number of departments– 196- randomly generated between 7 and 16 for every faculty;
- Number of lecturers– 1387- randomly generated between 4 and 10 for every department.

Activity	Execution Count		Average time				Total time			
	B/W/R	P	B	W	R	P	B	W	R	P
Prepare request for every faculty	1	1	5	100	38	21	5	100	38	21
Preview Request	16	16	1	2	1	9	9	44	16	155
Extract Data	1	16	1	25	13	9	1	25	13	156
Prepare request for every department	15	0	5	13	6	0	85	195	93	0
Prepare request for every lecturer	183	0	35	64	46	0	7830	11746	8524	0
Prepare personal report	1293	0	71	78	76	0	92011	101394	99399	0
Prepare department report	183	0	87	361	233	0	16095	66164	42696	0
Prepare faculty report	16	16	283	568	467	3	4542	9091	7487	51
Prepare university report	1	1	400	720	610	480	400	720	610	480

Table 1. Simulation results for every activity

It is important to notice that all the subprocesses are multi-instanced and asynchronous, i.e. all faculties prepare their reports at the same time, and all departments and lecturers are working the same way. That is why total estimations cannot be calculated by applying simple arithmetic operations to the input

variables. Moreover, this timing takes into account the actual work done.

The total time evaluation metric of the "cycle time" performance indicator is a subject of the greatest importance. This "*simple number*" allows measuring how much time is spent in preparing a single report. Accordingly, the "cycle time" values tell us that productivity and workforce turnover KPIs can be improved by implementing "The Authors" information system in each one of the faculties of the university. **Table 2** displays the simulation results for every subprocess as follows.

Subprocess	Execution Count		Average time				Total time			
	B/W/R	P	B	W	R	P	B	W	R	P
Faculty Data Gather	16	16	210	534	286	23	3368	8556	4581	379
Department Data Gather	183	0	150	428	160	0	26141	78499	29284	0
Lecturer Data Gather	1293	0	71	78	76	0	92028	101399	99408	0

Table 2. Simulation results for every subprocess

The obtained simulation results allow defining a realistic period for completing the business process. For instance, a Faculty would need between 210 and 534 minutes to submit the required data for the Preparation of the annual scientific report. On the other side, we can expect this activity to complete for about 286 minutes. This analysis also shows that the usage of information systems in the business process model may shorten the time for completion of this activity to 23 minutes. **Table 2** shows also that the information bottleneck is collection of data from individual lecturers and processing this information in the respective departments in case no information system is being used to store and summarize the thus collected data.

6 Conclusion

An approach for static and dynamic analysis of a business process model has been presented in this paper. Unlike other approaches the here proposed approach takes into consideration all the business processes comprising the business process model of the activities in enterprise. A custom developed software tool allows identifying deficiencies in the model and making management decisions for improving the overall performance of the business process activities in accordance with the in the set of KPIs for Operational performance. The proposed approach implementation is illustrated by means of a realistic case study of business process model offering services in an academic environment. Such processes are not easy to analyze, model, optimize and simulate because the execution of the business

process activities belonging to different business processes is interrelated and some of these activities execute in parallel. The results obtained by means of the proposed approach allow to evaluate both the logical and syntax correctness of the business processes comprising the model, as well as,the scope of documents, roles and information systems involved in the business processes. On the other side, the dynamic analysis allows using simulation for the evaluation of important metrics providing information how well the business process model matches the KPIs of the organization. .

Acknowledgement

This work was supported by the European Social Fund through the Human Resource Development Operational Program under contract BG051PO001-3.3.06-0052 (2012/2014)

References

1. **Giaglis, George M.** *A Taxonomy of Business Process Modeling and Information Systems Modeling Techniques*: International Journal of Flexible Manufacturing Systems, April 2001. pp. 209–228. Vols. 13, Issue 2.
2. **Jakob Freund, Bernd Rücker.** *Real-Life BPMN*: Camunda, 2012.
3. **Mendling, Jan.** *Metrics for Process Models.* Springer, 2008.
4. **Kristiyan Shahinyan, Evgeniy Krastev.** *Evaluation metrics for Business Processes in an Academic Environment.* Proceedings of the 7th International Conference on Information Systems & Grid Technologies, 2013. pp. 297–306.
5. **Rademakers, Tijs.** *Activiti in Action.* Manning, 2012.
6. **Kristiyan Shahinyan, Evgeniy Krastev.** *Extending a BPMN Engine with Evaluation Metrics for KPIs.* Proceedings of the Doctoral Conference in Mathematics, Informatics and Education, 2013. pp. 102- 109.
7. **Jan-Philipp Friedenstab, Christian Janieschy et. al.** *Extending BPMN for Business Activity Monitoring.* Proc. of the 45th ICSS , 2012.

Distributed Training and Testing Grid Infrastructure Evolution

Radoslava Hristova[1], Nikolay Kutovskiy[2,3], Vladimir Dimitrov[1], Vladimir Korenkov[2]

[1] Faculty of Mathematics and Informatics, University of Sofia "St. Kliment Ohridski", James Bourchier 5, 1164 Sofia, Bulgaria

[2] Laboratory of Information Technologies, Joint Institute for Nuclear Research, Joliot-Curie 6, 141980 Dubna, Moscow region, Russia

[3] National Scientific and Educational Centre of Particle and High Energy Physics of the Belarusian State University, Minsk, Belarus

radoslava@fmi.uni-sofia.bg, nikolay.kutovskiy@jinr.ru, cht@fmi.uni-sofia.bg, korenkov@jinr.ru

Abstract. The distributed training and testing grid infrastructure (t-infrastructure) continues developing. A set of deployed testbeds as well as a structure of some of them was changed. So a new testbed based on DIRAC middleware was setup to work on several activities for distributed computing of BES-III experiment. A virtual organization for NICA project was created and there is ongoing work on deploying few testbeds with different middleware to choose which one matches better the experiments' needs. Apart from that a set of grid sites integrated into EMI-based testbed was changed. Ongoing activities and future plans are covered as well.

1 Introduction

Grid technologies have already become a standard tool used by scientists in different fields. Moreover, commercial companies started to use grid for own needs and promote it for customers. To satisfy a growing demand in mastering of grid technologies, categories of different type of specialists need to be trained. Specialists like system administrators of grid sites, developers of grid services and applications, and end-users. There is a growing demand in such specialists who could use grid, deploy and maintain grid sites and services as well as develop new components and applications, which to run in grid environments. Unfortunately, the production grid infrastructures cannot be used for trainings, for middleware development or testing, because the common user has limited privileges, especially as regards middleware development and deployment.

In order to answer all of these requirements, a dedicated distributed training and testing grid infrastructure (t-infrastructure) was deployed at Joint Institute for Nuclear Research (JINR) [2]. One of the main goals of this infrastructure is to help to spread out knowledge and skills in grid technologies across scientists,

V. Dimitrov, V. Georgiev (Editors): ISGT'2014. ISSN 1314-4855
Proceedings of the 8th International Conference on
Information Systems and Grid Technologies, Sofia, 30–31. May, 2014.

researchers and students in educational and scientific organizations of JINR member states, one of which is Bulgaria [3].

In this article we observe the evolution of the t-infrastructure according to the changes of the underling middleware and related with it activities.

2 T-infrastructure Activities

Grid services of t-infrastructure are running inside virtual machines (VMs). This is permissible, because the performance of grid services for testing and training tasks is not a critical issue. The schema of EMI-based sites of t-infrastructure are shown on Figure 1.

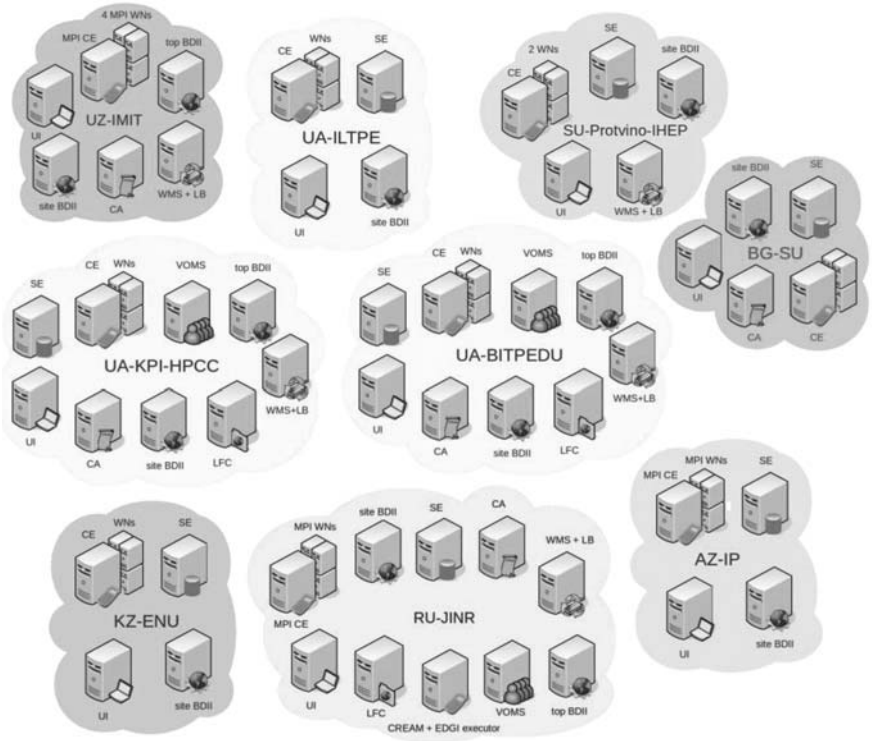

Fig. 1. EMI-based part of t-infrastructure

The educational site of the University of Sofia "St. Kliment Ohridski" (BG-SU) was initially added to the infrastructure in 2012. All deployed at that time grid services were gLite based ones. They had been upgraded to EMI in 2014. Currently, there are 4 grid services deployed on VMs running on 2 physical servers: CREAM, DPM SE, WN and UI. The certification authority is installed on physical machine.

The educational t-infrastructure was presented during the exercises of the course “Grid and cloud” to the students of the master program “Information systems” and the master program “Information-technology services and projects”. The grid technologies were introduced to more than 50 students.

The t-infrastructure is permanently used for giving trainings for students of University Centre of JINR, University “Dubna”, colleagues from JINR and also its member states. The trainings for system administrators are performed in a several ways: in person (trainees come to JINR or trainer come to organization where the training needs to be given) or remotely using some audio/videoconferencing tools.

The training for system administrators assumes dedicated resources with super-user privileges for each trainee. Trained administrators deploy grid sites at home organizations which can become a part of one of the global grid infrastructure (if grid site matches a set of requirements such as internet connectivity, computational and storage resources, ability to provide a required rate of reliability and availability) or can be integrated into t-infrastructure (if the grid site doesn’t match at least one of the mentioned requirements and hence can’t be a part of global grid infrastructure since t-infrastructure doesn’t have such strict requirements). As soon as grid site starts matching all requirements it can be reconfigured to become a part of the global grid infrastructure. There are no regular trainings for developers of grid services and grid-enabled applications yet. They are trained upon necessity for particular project.

3 T-Infrastructure Evolution

The t-infrastructure continues developing. All VMs with t-infrastructure services were migrated into the local private cloud [4]. A set of deployed testbeds as well as a structure of some of them were changed. So a new testbed based on DIRAC middleware was setup to work on several activities for distributed computing of BES-III experiment. A virtual organization for NICA project was created and there is ongoing work on deploying few testbeds with different middleware to choose which one matches better the experiments’ needs. Apart from that a set of grid sites integrated into EMI-based testbed was changed.

All components of t-infrastructure deployed at JINR cloud are shown on Fig. 2.

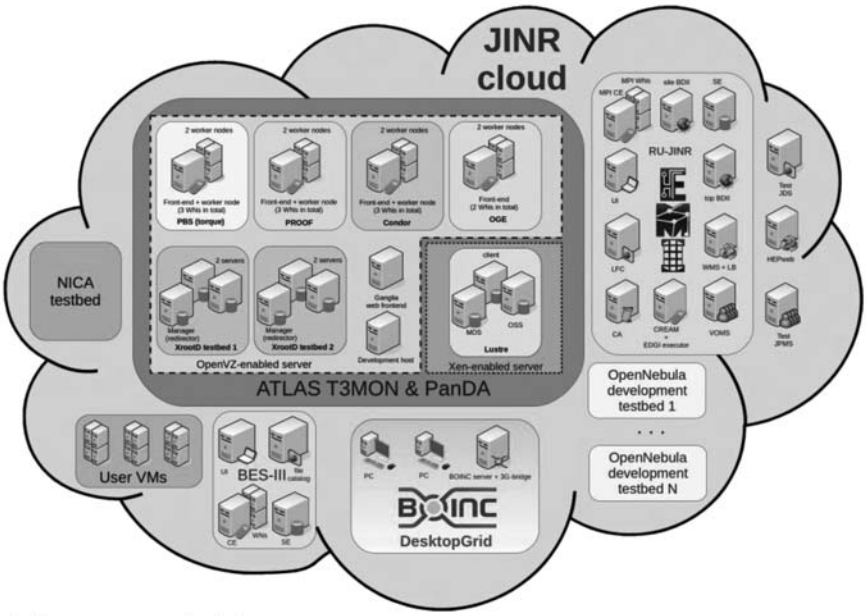

Fig. 2. Components of t-infrastructure deployed at JINR cloud

4 Conclusions

The t-infrastructure is successfully used for more than seven years for different tasks. It becomes a platform for training, research, development, tests and evaluation of modern technologies for distributed computing and data management. A set of t-infrastructure components are dynamically changed depending on current demands and priorities. T-infrastructure migration into the cloud helped to increase hardware resources utilization as well as significantly simply the job of system administrators by automating most of the virtual machines management tasks and by giving the users the ability to create and manage own VMs by themselves within the limit of the granted quotas.

Acknowledgments. The work reported in this paper is supported by the project Development of Grid Technologies at JINR and SU "St. Kliment Ohridski" - Information, Computer and Network Support of JINR activities", № 05-6-1048-2003/2013 and partially supported by the AYSS JINR grant #14-602-04.

References

1. Korenkov, V.V., Kutovskiy, N. A. “Distributed Training and Testing Grid Infrastructure”, Proceedings of the 4rd International Conference “Distributed Computing and Grid-technologies in Science and Education” (GRID’2010), Dubna, 2010, pp.148-152
2. Kutovskiy, N. A. “Distributed Training and Testing Grid Infrastructure Evolution”, Proceedings of the 5th International Conference “Distributed Computing and Grid-technologies in Science and Education” (GRID’2012), Dubna, 2012, pp.180-185
3. Kutovskiy, N. A. “Educational, Training and Testing Grid Infrastructure”, Proceedings of the XIV Conference of Young Scientists and Specialists (OMUS’ 2010), Dubna, 2010, pp.70-73
4. N. Balashov, A. Baranov, N. Kutovskiy, R. Semenov “Cloud Technologies Application at JINR”, to appear in proceedings of the 8th conference “Information Systems and Grid Technologies, Sofia, Bulgaria, 2014.

Author Index

Airchinnigh, Mícheál Mac an 49

Angeloska-Dichovska, Monika 93

Blazekovic, Marina 93

Balashov, Nikita 32

Baranov, Alexandr 32

Dimeski, Branko 14

Dimitrov, Vladimir 58, 115, 158

Georgiev, Vasil 65, 100

Grigorova, Katalina 25

Hasan, Falak 81

Hristov, Hristo 100

Hristova, Radoslava 115, 158

Ilchev, Velko 74

Kamenarov, Ivaylo 25

Kutovskiy, Nikolay 32, 158

Korenkov, Vladidmir 158

Krachunov, Milko 38

Krastev, Evgeniy 122, 144

Kulev, Ognyan 105

Nikolov, Vassil 49

Nikolov, Krasimir 133

Nisheva, Maria 74

Ranchev, Nikola 74

Savoska, Snezana 14

Semerdzhieva, Maria 122

Shahinyan, Kristiyan 144

Semenov, Roman 132

Simeonov, Daniel 65

Stojkovski, Viktorija 7, 93

Vasileva, Svetlana 133

Vassilev, Dimitar 74

Veljanovska, Kostandina 7