GENERAL STRUCTURE OF COLLECT/REPORT PARADIGM FOR STORING AND ACCESSING BIG DATA

Krassimir Markov, Krassimira Ivanova

Abstract: *The general structure of Collect/Report Paradigm (CRP) is outlined in this paper. CRP is a new approach for storing and accessing large amounts of data. Main advantages of CRP are: (1) collecting information is done by all nodes independently in parallel. It is possible one node to send information to another; (2) reporting information is provided only by the nodes which really contain information related to the request; the rest nodes do not react, they remain silent; (3) input data as well as results are in RDF-triple or RDF-quadruple format.*

Keywords*: Collect/Report Paradigm, Big Data, Cloud computing.*

ACM Classification Keywords: *D.4.3 File Systems Management, Access methods.*

Introduction

The Collect/Report Paradigm (CRP) is a new approach for storing and accessing large amounts of data.

There are several interesting areas of implementing of the Collect/Report Paradigm, for instance, business applications where flexibility of this approach will give some new possibilities; linguistic systems which work with large linguistic data sets; cognitive modeling; etc.

Maybe the most interesting is the area of so called "Big Data". The term Big Data applies to information that can't be processed or analyzed using traditional processes or tools. Increasingly, organizations today are facing more and more "Big Data challenges". They have access to a wealth of information, but they don't know how to get value out of it because it is sitting in its most raw form or in a semi-structured or unstructured format [Zikopoulos et al, 2012].

Three main characteristics define Big Data: ***Volume, Variety, and Velocity*** [Zikopoulos et al, 2012].

- **Volume** (the sheer volume of data being stored today is exploding) – avoiding additional indexing, duplication of keywords, and corresponded pointers, leads to reducing additional memory needed for accessing information i.e. we may use addressing but not classical search engines;
- **Velocity** (a conventional understanding of velocity typically considers how quickly the data is arriving and stored, and its associated rates of retrieval) – avoiding recompilation

of information base permits high speed of storing and immediately readiness of information to be accessed. This is very important possibility for stream data;

- **Variety** (it represents all types of data — a fundamental shift in analysis requirements from traditional structured data to include raw, semi-structured, and unstructured data as part of the decision-making and insight process) – natural language addressing permits creating a special kind of graph information bases which may operate both with structured as well as semi-structured information.

These characteristics cause corresponded problems of storing Big Data which may be solved by means of CRP.

Popular approach for representing Big Data is Resource Definition Framework (RDF). Let remember, RDF is a graph based data format which is schema-less, thus unstructured, and self-describing, meaning that graph labels within the graph describe the data itself. The prevalence of RDF data is due to variety of underlying graph based models, i.e. almost any type of data can be expressed in this format including relational and XML data [Faye et al, 2012].

Big Data created the need for a new class of capabilities to augment the way things are done today to provide better line of site and controls over our existing knowledge domains and the ability to act on them.

MapReduce Paradigm

In the Big Data community, the "MapReduce Paradigm" has been seen as one of the key enabling approaches for meeting the continuously increasing demands on computing resources imposed by massive data sets. MapReduce is a highly scalable programming paradigm capable of processing massive volumes of data by means of parallel execution on a large number of commodity computing nodes. It was recently popularized by Google [Dean & Ghemawat, 2008], but today the MapReduce paradigm has been implemented in many open source projects, the most prominent being the Apache Hadoop [Hadoop, 2014]. The popularity of MapReduce can be accredited to its high scalability, fault-tolerance, simplicity and independence from the programming language or the data storage system.

At the same time, MapReduce faces a number of obstacles when dealing with Big Data including the lack of a high-level language such as SQL, challenges in implementing iterative algorithms, support for iterative ad-hoc data exploration, and stream processing [Grolinger et al, 2014].

A possible solution may be the Collect/Report Paradigm. It is suitable for storing Big Data in large information bases located on different storage systems – from personal computers up to cloud servers.

Collect/Report Paradigm

The idea of Collect/Report Paradigm (CRP) is very simple and because of this it is perspective to be realized. Similar model one may see in the game of chance "Bingo" (Figure 1) for two or more players, who mark off numbers on a grid with unique sequence of numbers printed on their individual cards as they are announced by the Caller corresponding to numbered balls drawn at random; the game is won by the first person to call out "bingo!" or "house!" after crossing off all numbers on the grid or in one line of the grid [YourDictionary, 2013].

To play Bingo one has to "***collect***" (to buy) one or more individual cards and after starting the game to listen what number the Caller will announce, to find in the individual cards the same numbers and to mark them (i.e. to process the stream of incoming data). After marking every new number, (in real time, before next number will be announced) player has to analyze the configuration of marked cells on the individual cards and to decide if it is the winner configuration. If the configuration is a winner one, the player has to "***report***" (to call out) "Bingo". Only the players with winner configurations have to report, the others must stay silent.

Figure 1. Illustration of Collect/Report Paradigm via example of Bingo game

Collect/Report Paradigm is based on the possibility of so called "Natural Language Addressing" (NLA). [Ivanova et al, 2012a; Ivanova et al, 2012b; Ivanova et al, 2013b; Ivanova et al, 2013c; Ivanova et al, 2013e; Ivanova, 2014a; Ivanova, 2014b, Ivanova, 2014c; Ivanova, 2014d; Ivanova, 2014e; Ivanova, 2014f].

CRP assumes that incoming information is coded in RDF format. Using first element of triple as address and the second element as name of layer (archive), the third element may be stored and accessed directly. In other words, we have many different layers stored in separate archives which may be distributed all over the world. The correspondence between archives is strongly kept by names as addresses which are equal for all layers.

In Collect/Report Paradigm, all nodes have to "listen" in parallel the incoming stream of RDF-data and to "collect" (to store) information only in the layers the nodes have to support. In the same time, nodes have to "listen" incoming stream of requests and only nodes, which have information corresponded to given request has to "report" (to send answer).

As an example let's see Table 1. It represents six nodes numbered from 1 to 6 which may be distributed all over the net. Incoming information is in RDF triples (subject, relation, object). Information (objects) for the same subject and relation is concatenated in the corresponded points. Let assume that the Table 1 represents the state of nodes at given time moment. If in this moment a request for word "cut" will come, only nodes 1 and 6 will "report" the content (definitions) from corresponded cells. Node 1 will report only the first row which correspond to "cut" with small letters but not its second row which corresponds to word "CUT" with capital letters. Nodes 2, 3, 4, and 5 will rest silent.

Table 1. Content of six sample nodes

node	layer	NLA	definition
1	**adj_all**	cut	{ cut, shortened, (with parts removed; "the drastically cut film") } { cut, thinned, weakened, (mixed with water; "sold cut whiskey"; "a cup of thinned soup") } { cut, slashed, ((used of rates or prices) reduced usually sharply; "the slashed prices attracted buyers") } { cut, emasculated, gelded, ((of a male animal) having the testicles removed; "a cut horse") }

node	layer	NLA	definition
1	**adj_all**	CUT	{ [CUT1, UNCUT1,!] (separated into parts or laid open or penetrated with a sharp edge or instrument; "the cut surface was mottled"; "cut tobacco"; "blood from his cut forehead"; "bandages on her cut wrists") } { [CUT2, UNCUT2,!] ((of pages of a book) having the folds of the leaves trimmed or slit; "the cut pages of the book") } { [CUT3, UNCUT3,!] (fashioned or shaped by cutting; "a well-cut suit"; "cut diamonds"; "cut velvet") }
2	**adj_pert**	cut	empty definition
3	**adj_ppl**	cut	empty definition
4	**adv_all**	cut	empty definition
5	**noun_Tops**	cut	empty definition
6	**noun_act**	cut	{ cut6, absence,@ (an unexcused absence from class; "he was punished for taking too many cuts in his math class") } { cut5, reduction,@ (the act of reducing the amount or number; "the mayor proposed extensive cuts in the city budget") } { cut, [cutting, verb.creation:cut11,+] cutting_off1, shortening,@ (the act of shortening something by chopping off the ends; "the barber gave him a good cut") } { cut1, [cutting1, verb.contact:cut10,+ verb.contact:cut,+] division,@ (the act of cutting something into parts; "his cuts were skillful"; "his cutting of the cake made a terrible mess") } { cut2, [cutting2, verb.contact:cut10,+] opening2,@ (the act of penetrating or opening open with a sharp edge; "his cut in the lining revealed the hidden jewels") } { cut9, [cutting9, verb.contact:cut5,+] division,@ card_game,#p (the

node	layer	NLA	definition
			division of a deck of cards before dealing; "he insisted that we give him the last cut before every deal"; "the cutting of the cards soon became a ritual") } { cut8, [undercut, verb.contact:undercut,+] stroke,@ tennis,;c badminton,;c squash,;c ((sports) a stroke that puts reverse spin on the ball; "cuts do not bother a good tennis player") }

In general, Collect/Report Paradigm is illustrated on Figure 2.

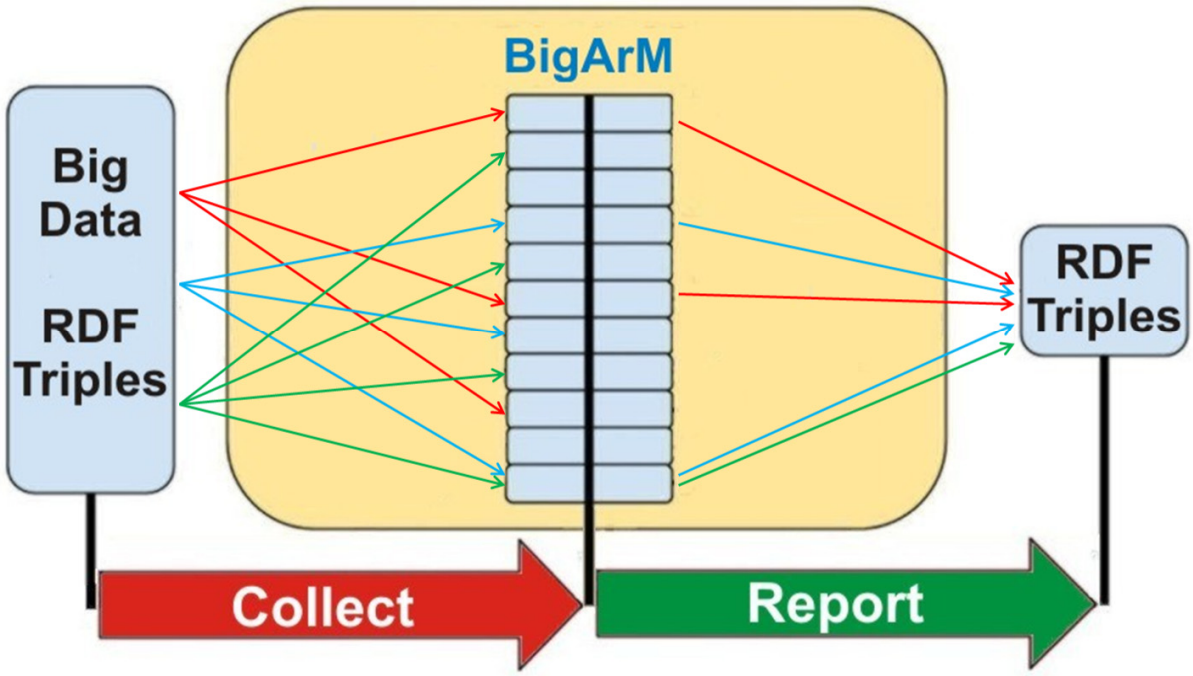

Figure 2. Cloud Collect/Report Scheme for Storing and Accessing Big Data

Main advantages of Collect/Report Paradigm (Figure 2) are:

- Collecting information is done by all nodes independently in parallel. It is possible one node to send information to another;
- Reporting information is provided only by the nodes which really contain information related to the request; the rest nodes do not react, they remain silent;
- Input data as well as results are in RDF-triple or RDF-quadruple format.

Collect/Report Paradigm has multi-level structure (Figure 3).

The first and second levels are "basic" and consist of two mathematical models:

- Multi-Domain Information Model [Markov, 2004];
- Natural Language Addressing Model [Ivanova et al, 2013a; Ivanova et al, 2013d].

The third and fourth levels are "methodical" and consist of two access methods:

- Multi-Domain Access Method [Markov, 1984];
- Natural Language Access Method [Ivanova, 2014g; Ivanova, 2014h].

Fifth level consists of program realizations of the access methods [Markov et al, 1990; Markov et al, 2008]. The ongoing new realization is called "BigArM".

Finally, the upper two levels are External data models (Graph Information Bases) and Operational Environment (Cloud).

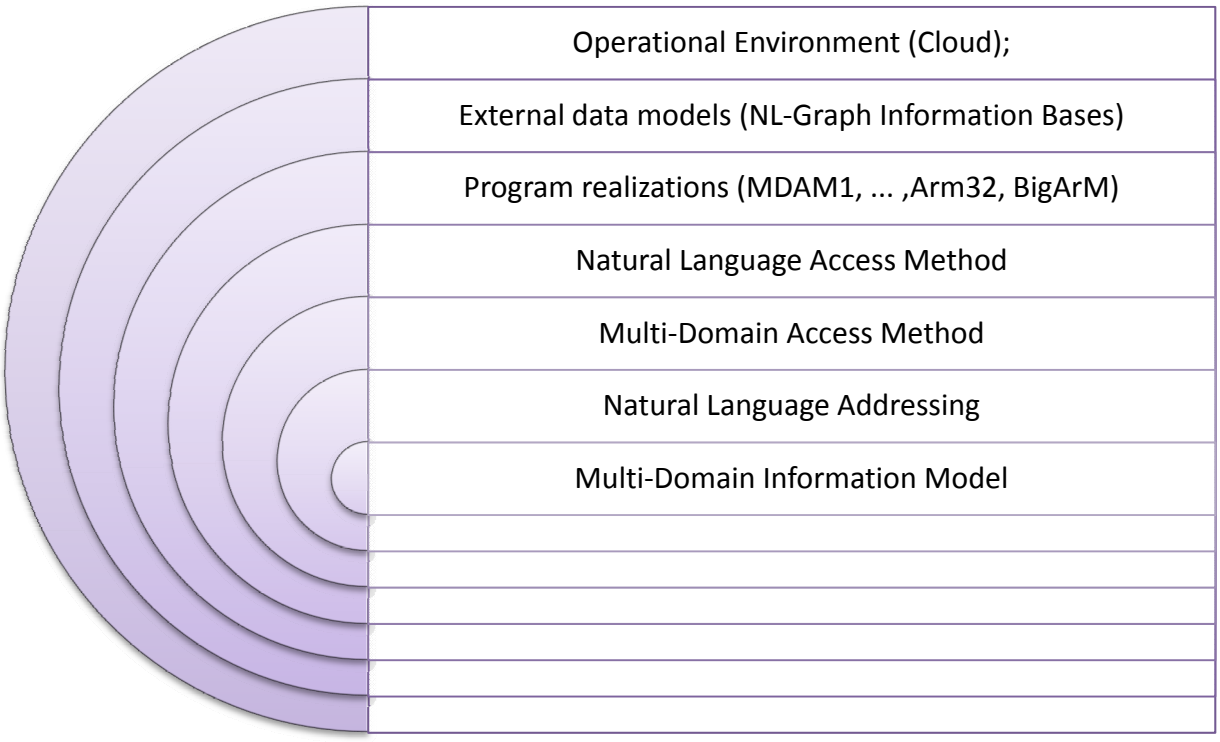

Figure 3. Main levels of the CRP structure

Conclusion

The milestone for the work presented in this paper is the simple idea that we may use a special kind of organization of the information and this way to develop easy to use information bases and with very high speed for response which enables ***the real-time analytical processing*** (RTAP) [Markov, 2005]. (The RTAP multithreaded processing engine needs to support extremely large volumes of data in real time. The analytics performed are composed of combinations of algorithmic, statistical and logical functions. [B-Jensen, 2002])

In the same time, CRP is good foundation for intelligent data processing based on multi-dimensional memory structures [Markov et al, 2013].

Finally, as it was mentioned in [Markov et al, 2014], via CRP and natural language addressing, three main problems of storing Big Data may be solved:

- Volume – avoiding additional indexing, duplication of keywords, and corresponded pointers, leads to reducing additional memory needed for accessing information i.e. we may use addressing but not classical search engines;
- Velocity – avoiding recompilation of information base permits high speed of storing and immediately readiness of information to be accessed. This is very important possibility for stream data;
- Variety – natural language addressing permits creating a special kind of graph information bases which may operate both with structured as well as semi-structured information.

Bibliography

[B-Jensen, 2002] M.T. B-Jensen. High Tower Software's Tower View is the Odds-On Favorite of International Game Technology for Real-Time Data Management. Product Review published in DM Review Magazine July 2002 Issue. http://www.dmreview.com/article_sub.cfm?articleId=5403

[Dean & Ghemawat, 2008] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, 51(1), 2008, pp. 107-113.

[Faye et al, 2012] David C. Faye, Olivier Cure, Guillaume Blin, "A survey of RDF storage approaches", Received, December 12, 2011, Accepted, February 7, 2012, ARIMA Journal, vol. 15, 2012 pp. 11-35.

[Grolinger et al, 2014] K. Grolinger, M. Hayes, W. Higashino, A. L'Heureux, D. S. Allison, M. A. M. Capretz, "Challenges for MapReduce in Big Data", Proc. of the IEEE 10th 2014 World Congress on Services (SERVICES 2014), Alaska, USA, June 27-July 2, 2014

[Hadoop, 2014] Apache Hadoop, http://hadoop.apache.org . (accessed: 22.12.14)

[Ivanova et al, 2012a] Krassimira Ivanova, Vitalii Velychko, Krassimir Markov, "About NLaddressing", (К вопросу о естествено-языконой адрессации) In: V. Velychko et al (ed.), Problems of Computer in Intellectualization. ITHEA® 2012, Kiev, Ukraine - Sofia, Bulgaria, ISBN: 978-954-16-0061-0 (printed), ISBN: 978-954-16-0062-7 (online), pp. 77-83 (in Russian).

[Ivanova et al, 2012b] Krassimira Ivanova, Vitalii Velychko, Krassimir Markov, "Storing RDF Graphs using NL-addressing", In: G. Setlak, M. Alexandrov, K. Markov (ed.), Artificial Intelligence Methods and Techniques for Business and Engineering Applications. ITHEA® 2012, Rzeszow, Poland; Sofia, Bulgaria, ISBN: 978-954-16-0057-3 (printed), ISBN: 978-954-16-0058-0 (online), pp. 84 – 98.

[Ivanova et al, 2013a] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "Introduction to the Natural Language Addressing", International Journal "Information Technologies & Knowledge" Vol.7, Number 2, 2013, ISSN 1313-0455 (printed), 1313-048X (online), pp. 139–146.

[Ivanova et al, 2013b] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "Introduction to Storing Graphs by NL-Addressing", International Journal "Information Theories and Applications", Vol. 20, Number 3, 2013, ISSN 1310-0513 (printed), 1313-0463 (online), pp. 263 – 284.

[Ivanova et al, 2013c] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "Storing Dictionaries and Thesauruses Using NL-Addressing", International Journal "Information Models and Analyses" Vol.2, Number 3, 2013, ISSN 1314-6416 (printed), 1314-6432(online), pp. 239 - 251.

[Ivanova et al, 2013d] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "The Natural Language Addressing Approach", International Scientific Conference "Modern Informatics: Problems, Achievements, and Prospects of Development", devoted to the 90th anniversary of academician V. M. Glushkov. Kiev, Ukraine, 2013, ISBN 978-966-02-6928-6, pp. 214 - 215.

[Ivanova et al, 2013e] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "Storing Ontologies by NL-Addressing", IVth All–Russian Conference "Knowledge-Ontology-Theory" (KONT-13), Novosibirsk, Russia, 2013, ISSN 0568-661X, pp. 175 - 184.

[Ivanova, 2014a] Kr. Ivanova, "Storing Data using Natural Language Addressing", PhD Thesis, Hasselt University, Belgium, 2014, 340 p.

[Ivanova, 2014b] Krassimira Ivanova, „A Solution of Three Main Problems of Big Data", 4-я Международная конференция "Управление знаниями и конкурентная разведка" в рамках 18-го Юбилейного Международного молодежного форума "Радиоэлектроника и молодежь в XXI веке", том 9, Харьков, Украйна, 2014, pp. 6-9.

[Ivanova, 2014c] Krassimira Ivanova, "Example of Multi-Layer Knowledge Representation by means of Natural Language Addressing", In: V. Velychko, O. Voloshyn, K. Markov, (eds.), proceedings of the

XX-th International Conference "Knowledge-Dialogue-Solution", ITHEA®, Kyiv, Ukraine, Sofia, Bulgaria, 2014, ISSN 1313-0087 (printed), ISSN 1313-1206 (online), pp. 115 - 117.

[Ivanova, 2014d] Krassimira Ivanova, "WORDArM - A System for Storing Dictionaries and Thesauruses by Natural Language Addressing", International Journal "Information Theories and Applications", Vol. 21, Number 4, 2014, ISSN 1310-0513 (printed), 1313-0463 (online), pp. 362 - 370.

[Ivanova, 2014e] Krassimira Ivanova, "ONTOArM - a System for Storing Ontologies by Natural Language Addressing", International Journal "Information Technologies & Knowledge", Vol. 8, Number 4, 2014, ISSN 1313-0455 (printed), 1313-048X (online), pp. 303 - 312.

[Ivanova, 2014f] Krassimira Ivanova, "RDFArM - A System for Storing Large Sets of RDF Triples and Quadruples by means of Natural Language Addressing", International Journal "Information Models and Analyses", Vol. 3, Number 4, 2014, ISSN 1314-6416 (printed), 1314-6432 (online), pp. 303 - 322.

[Ivanova, 2014g] Krassimira Ivanova, "Multi-Layer Knowledge Representation", International Journal "Information Content and Processing", Vol. 1, Number 4, 2014, ISSN 2367-5128 (printed), 2367-5152 (online), pp. 303 - 310.

[Ivanova, 2014h] Krassimira Ivanova, "Practical Aspects of Natural Language Addressing", In: G. Setlak, K. Markov (ed.), Computational Models for Business and Engineering Domains, ITHEA®, 2014, Rzeszow, Poland, Sofia, Bulgaria, ISBN: 978-954-16-0066-5 (printed), ISBN: 978-954-16-0067-2 (online), pp. 172 – 186.

[Markov et al, 1990] K. Markov, T. Todorov, V. Nikolov, "Multidomain Access Method for the IBM PC", Research in Informatics, Vol. 3, Academie-Verlag Berlin, 1990, pp. 218-230.

[Markov et al, 2008] Markov, K., Ivanova, K., Mitov, I., & Karastanev, S., "Advance of the access methods", International Journal of Information Technologies and Knowledge, 2(2), 2008, pp. 123–135.

[Markov et al, 2013] Markov, Krassimir, Koen Vanhoof, Iliya Mitov, Benoit Depaire, Krassimira Ivanova, Vitalii Velychko and Victor Gladun, "Intelligent Data Processing Based on Multi- Dimensional Numbered Memory Structures", Diagnostic Test Approaches to Machine Learning and Commonsense Reasoning Systems, IGI Global, 2013, pp. 156-184, doi:10.4018/978-1-4666-1900-5.ch007, ISBN: 978 1-4666-1900-5, EISBN: 978-1-4666- 1901-2
Reprinted in: Markov, Krassimir, Koen Vanhoof, Iliya Mitov, Benoit Depaire, Krassimira Ivanova, Vitalii Velychko and Victor Gladun, "Intelligent Data Processing Based on Multi-Dimensional Numbered Memory Structures", Data Mining: Concepts, Methodologies, Tools, and Applications, IGI Global, 2013, pp. 445-473, doi:10.4018/978-1-4666-2455-9.ch022, ISBN13: 978-1-4666-2455-9, EISBN13: 978-1-4666-2456-6

[Markov et al, 2014] Kr. Markov, Kr. Ivanova, K. Vanhoof, B. Depaire, V. Velychko, J. Castellanos, L. Aslanyan, St. Karastanev, "Storing Big Data Using Natural Language Addressing", In: N. Lyutov (ed.), Int. Sc. Conference "Informatics in the Scientific Knowledge", VFU, Varna, Bulgaria, 2014, ISSN: 1313-4345, pp. 147-164.

[Markov, 1984] Markov Kr., "A Multi-domain Access Method", Proceedings of the International Conference on Computer Based Scientific Research, Plovdiv, 1984, pp. 558-563.

[Markov, 2004] Markov, K., "Multi-domain information model", Int. J. Information Theories and Applications, 11/4, 2004, pp. 303-308

[Markov, 2005] Markov, K., "Building data warehouses using numbered multidimensional information spaces", International Journal of Information Theories and Applications, 12(2), 2005, pp. 193–199

[YourDictionary, 2013] YourDictionary, "LoveToKnow", http://www.yourdictionary.com (accessed: 20.07.2013).

[Zikopoulos et al, 2012] Paul C. Zikopoulos, Chris Eaton, Dirk de Roos, Thomas Deutsch, George Lapis, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", Copyright© 2012 by The McGraw-Hill Companies, ISBN 978-0-07-179053-6, MHID 0-07-179053-5, 2012, 166 p.

Authors' Information

Krassimir Markov *– Institute of Mathematics and Informatics at Bulgarian Academy of Sciences; Bulgaria; e-mail: markov@foibg.com*

Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems

Ivanova Krassimira *– University of National and World Economy, Sofia, Bulgaria; Institute of Mathematics and Informatics, BAS, Bulgaria; e-mail: krasy78@mail.bg*

Major Fields of Scientific Research: Software Engineering, Business Informatics, Data Mining, Multidimensional multi-layer data structures in self-structured systems