Accepted Manuscript

Third International Workshop on Modeling and Management of Big Data (MoBiD'14)

David Gil, Il-Yeol Song

PII: S0167-739X(15)00251-4
DOI: http://dx.doi.org/10.1016/j.future.2015.07.019
Reference: FUTURE 2810

To appear in: *Future Generation Computer Systems*

Received date: 14 June 2015
Revised date: 28 July 2015
Accepted date: 31 July 2015

Please cite this article as: D. Gil, I.-Y. Song, Third International Workshop on Modeling and Management of Big Data (MoBiD'14), *Future Generation Computer Systems* (2015), http://dx.doi.org/10.1016/j.future.2015.07.019

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Editorial
Third International Workshop on Modeling and Management of Big Data (MoBiD'14)

David Gil[a], Il-Yeol Song[b]

[a]*Lucentia Research Group, Computing Technology and Data Processing, University of Alicante, Spain*
[b]*College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, U.S.A.*

The aim of the International Workshop on Modeling and Management of Big Data is to bring together researchers, developers and practitioners to discuss research issues and experience in modeling, developing and deploying systems and techniques to deal with Big Data. The third International Workshop on Modeling and Management of Big Data (MoBiD'14), held in Atlanta, October, 27-30, 2014 was a continuation/evolution of the previous workshops, the International Workshop on Modeling for Data-Intensive Computing (MoDIC'12), held in Florence, Italy, October 15-18, 2012 and MoBiD'13 held in Hong Kong, November 11-13, 2013. MoBiD'14 was presented with the aim to attract papers on the latest and best proposals for modeling and managing Big Data in this new era of the data-drive paradigm. This new conceptualization of big data applications incorporating both internal and external Big Data requires new models and methods to accomplish their conceptual modeling phase. Papers focusing on the application and the use of conceptual modeling approaches for Big Data, MapReduce, Hadoop and its ecosystems, Big Data Analytics, social networking, cloud computing, security and privacy, data science, etc. were highly encouraged. Therefore, the workshop has been an international forum for researchers and practitioners who are interested in the different facets related to the use of the conceptual modeling approaches for the development of next generation of applications based on Big Data. We view that several key themes with the Big Data trend include (i) using a cloud for large-scale external and internal data; (ii) providing an easy-to-use but powerful services to access/manage/analyze the big data in the cloud; (iii) defining a problem-solving space and developing an architecture for a big data environment to conceptualize goals, tasks, and problem-solving methods to apply to domains; and (iv) managing big data and analyzing them to discover business values.

MoBiD'14 attracted papers from 9 different countries distributed all over the world: France, Greece, India, Japan, Kenya, Korea, Spain, United Kingdom and USA. We have finally received 14 papers and the Program Committee has selected 5 papers, making an acceptance rate of 35%. In the following, we summarize these selected papers:

ACCEPTED MANUSCRIPT

The first paper, "From Business Intelligence to Semantic Data Stream Management" by Marie-Aude Aufaure and Raja Chiky [3], introduce recent work on Real-Time Business Intelligence that utilizes semantic data stream management. This paper addresses the new tendencies of real-time systems that are continuously generating data to be analyzed, processed, and stored. They also present underlying approaches to continuous queries and data summarization.

The second paper, entitled "Business Intelligence and Big Data in the Cloud: Opportunities for Design-Science Researchers" by Odette Sangupamba Mwilu, Nicolas Prat and Isabelle Comyn-Wattiau [25] deals with the new opportunities for business intelligence (BI) and analytics offered by Cloud computing and big data. They propose a typology of artifacts potentially produced by researchers in design science. Then, after analyzing the state of the art through that typology, they use it to sketch opportunities of new research to improve BI and analytics capabilities in the cloud and from big data.

The third paper, "A Data Quality in Use Model for Big Data" by Ismael Caballero, Bibiano Rivas, Manuel Serrano and Mario Piattini [5] is a position paper that proposes the 3Cs model, which is composed of three data quality dimensions for assessing the quality-in-use of big datasets: Contextual Consistency, Operational Consistency and Temporal Consistency. The aim is that the quality of data lacks a quality-in-use model adapted for big data.

The fourth paper, "Energy consumption prediction by using an integrated multidimensional modeling approach and data mining techniques with Big Data" by Jess Peral, Antonio Ferrndez, Roberto Tardo, Alejandro Mat, Elisa de Gregorio [24] explores the opportunities of using ICT (Information and Communication Technologies) as an enabling technology to reduce energy consumption in cities. It proposes a multidimensional hybrid architecture that makes use of current energy data and external information (with unstructured data sources) to improve knowledge acquisition and allow managers to make better decisions.

The last paper of this Special Issue, entitled "Benchmarking Performance for Migrating a Relational Application to a Parallel Implementation" by Krishna Karthik Gadiraju, Karen C. Davis, and Paul G. Talaga [15] investigates the impact of scaling up the data sizes for several benchmarking queries. They illustrate what kind of performance results an organization could expect when they migrate current applications to big data environments. The authors measure the speedup for query execution for all dataset sizes resulting from the scale up. They conclude that Hive loads the large datasets faster than MySQL, while it is marginally slower than MySQL when loading the smaller datasets.

Finally, we would like to thank all the authors who revised and extended their papers for this special issue and the reviewers for their hard work in revising these extended papers twice and providing critical and useful comments that helped the authors in improving their papers. Absolutely, all of them have contributed to creating this special issue of a high quality. We hope the readers will enjoy reading this issue and find the content beneficial to their research.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

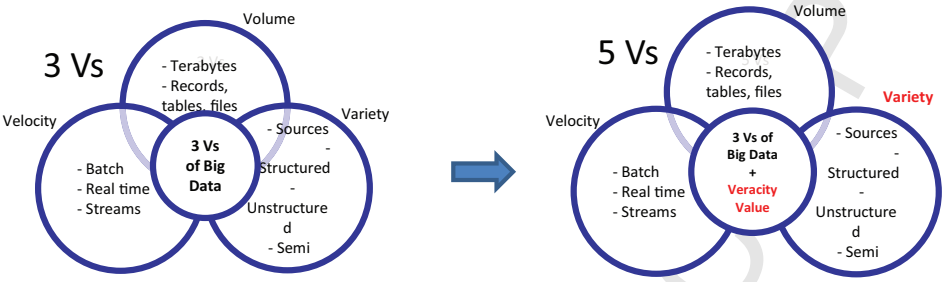

Figure 1: Defining big data with 3 V's and moving towards 5 V's

Conceptual modeling in the big data era

The experience in conjunction with the novelty and the new trends developed during the last three years lead us to summarize our thoughts and expectation in these topics for this and the next editions of this workshop.

Big data is a very broad term which is often easily understood by means of a graphical representation (Fig. 1) in order to not only pay attention to the "Big" word, but especially to understand that "big data" express the difficulty into dealing with data in different dimensions.

Currently there are too many scenarios where the term Big Data appears. Scientists, business executives, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics [9] [22].

There are many domains whose data management needs have exploded. For example, we can discuss data management challenges of E-commerce along the three dimensions: volume, velocity and variety.

- On Volume: "The lower cost of e-channels enables an enterprise to offer its goods or services to more individuals or trading partners. The explosion of the data to be collected in e-commerce are even up to 10x of the quantity of data about an individual transaction, thereby significantly increasing the overall volume of data to be managed."
- On Velocity: "E-commerce has also increased point-of-interaction (POI) speed, and consequently the pace data used to support interactions and generated by interactions."
- On Variety: "No greater barrier to effective data management will exist than the variety of incompatible data formats, non-aligned data structures, and inconsistent data semantics."

Where does big data come from? (i) "data exhaust" from customers; (ii) new and pervasive sensors; (iii) the ability to "keep everything" [23] [13].

ACCEPTED MANUSCRIPT

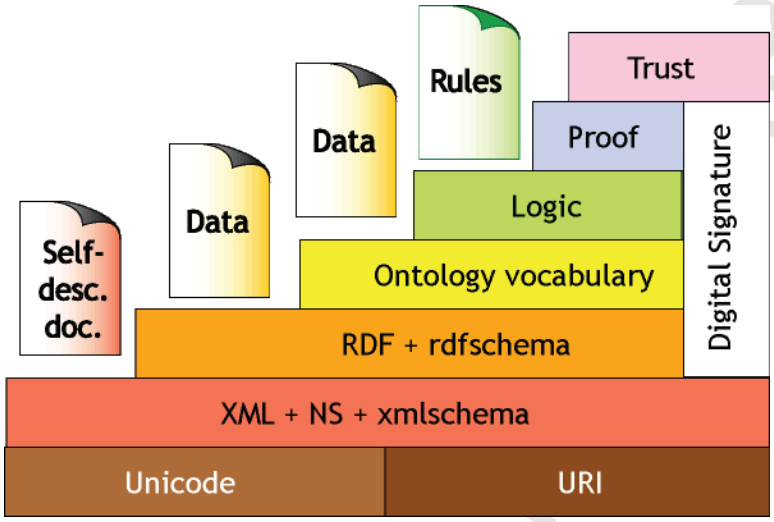

Figure 2: Conceptualization

In [4] it is indicated that with the significant advances in Information and Communications Technology (ICT) over the last half century, there is an increasingly perceived vision that computing will one day be the 5th utility (after water, electricity, gas, and telephony). It is defined Cloud computing and provide the architecture for creating Clouds with market-oriented resource allocation by leveraging technologies such as Virtual Machines (VMs). The proliferation of the devices in a communicatingactuating network creates the Internet of Things (IoT) [17]. In [11] it is analysed the challenges and requirements for next-generation Big Data services and presented a solution designed to support next-generation Big Data applications.

Regarding the difficulty of managing Big Data, it has been stated that Big Data is any data that is expensive to manage and hard to extract value from [14]. Among the Vs shown in Fig. 1, in this brief summary, we will focus on The V of Variety. This will lead us to the main topic of the workshop which is Conceptual modeling of Big Data.

With the help of various conceptual modeling techniques, such as ontologies (Web Ontology Language -OWL), semantic, RDF schemas, SPARQL Language, etc, it could be possible to formulate novel integration architectures. The framework will make it possible to take advantage of data integration on the Web [10]. Embley explains in his paper entitled “Big Data Conceptual Modeling to the Rescue” [12] several layers of this conceptualization including examples as well as various techniques for this goal (Fig. 2).

The intersection of the terms “Conceptual Modeling” & “BIG DATA” still appear as one of the challenges in big data. We keep finding Volume too big, Variety too many, Velocity too fast, and Veracity too uncertain.

ACCEPTED MANUSCRIPT

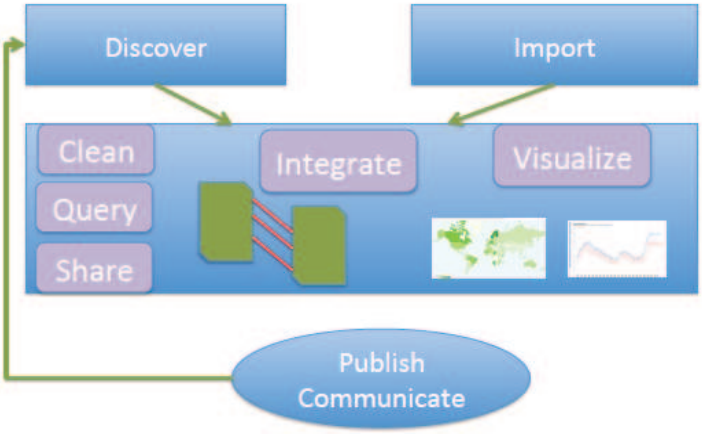

Figure 3: The goal is a Structured Data Ecosystem

Chen's article [8] illustrates to show historical groundings with comments about its original contributions (including mapping to database) - to organize data well for search and business transaction processing.

For instance, the conceptualization of the Web includes semantic search as well as keyword search and World-wide knowledge sharing. There are some examples, such as DB-pedia, Conceptual Graphs (like Google's Knowledge Graph, Yahoo!'s Web of Object, Facebook's Graph Search, Microsoft's/Bing's Satori Knowledge Base), Metaweb, FamilySearch.

The World-Wide Web provides access to millions of data tables with high-quality content, formatted either in HTML tables, HTML lists, or other structured formats, or stored in on-line data management services. These tables contain data about virtually every domain of interest to mankind. Several research projects aim at enabling search over these data sets and ultimately the ability to answer queries and to combine data from multiple sources. In this context one of the main goal is to achieve structured Data in an Ecosystem as shown Fig. 3.

Halevy and others researches have contribute enormously to the search of the structured data on the Web, integration [18] [19] [7] [6]. There exist some tools like using fusion tables as they are easy-to-use. They are database systems that are integrated with the Web [16].

Another idea is to use WebTables, which basically is discovering a (structured) needle in an (unstructured) haystack. The challenges are (i) Finding the good tables on the Web; (ii) Understanding their semantics; (iii) Understanding user's intentions ref The Needle in the Haystack is to find high quality HTML tables as Very often semantics embedded in surrounding text makes it harder.

Some mechanisms in order to solve this situation are:

ACCEPTED MANUSCRIPT

- The People's Ontology [Open Information Extraction]. Mine a database of entities and classes from the Web [2] [1]
- Recovering Table Semantics [26].
- Recovering Binary Relationships [26]
- Attribute Correlations [20]
- Synonym Discovery [21]

In conclusion, we can state that although there is a lot of work already done, there is much more to do. Therefore the challenges of incorporating the more and more researches from many different areas it has to keep increasing in order to be able to manage big data more efficiently in the near future.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

[1] H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, and N. R. Shadbolt. Automatic ontology-based knowledge extraction from web documents. *Intelligent Systems, IEEE*, 18(1):14–21, 2003.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[3] M.-A. Aufaure and R. Chiky. From business intelligence to semantic data stream management. *Submitted to Future Generation Computer Systems*.

[4] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic. Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems*, 25(6): 599–616, 2009.

[5] I. Caballero, B. Rivas, M. Serrano, and M. Piattini. A data quality in use model for big data. *Submitted to Future Generation Computer Systems*.

[6] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. *Proceedings of the VLDB Endowment*, 1(1):538–549, 2008.

[7] M. J. Cafarella, A. Halevy, and J. Madhavan. Structured data on the web. *Communications of the ACM*, 54(2):72–79, 2011.

[8] P. P.-S. Chen. The entity-relationship modeltoward a unified view of data. *ACM Transactions on Database Systems (TODS)*, 1(1):9–36, 1976.

[9] K. Cukier. *Data, data everywhere: A special report on managing information*. Economist Newspaper, 2010.

[10] A. Doan, A. Halevy, and Z. Ives. *Principles of data integration*. Elsevier, 2012.

[11] C. Dobre and F. Xhafa. Intelligent services for big data science. *Future Generation Computer Systems*, 37:267–281, 2014.

[12] D. W. Embley and S. W. Liddle. Big dataconceptual modeling to the rescue. In *Conceptual Modeling*, pages 1–8. Springer, 2013.

[13] W. Fan and A. Bifet. Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*, 14(2):1–5, 2013.

[14] M. J. Franklin. Making sense of big data with the berkeley data analytics stack. In *SSDBM*, page 1, 2013.

[15] K. K. Gadiraju, K. C. Davis, and P. G. Talaga. Benchmarking performance for migrating a relational application to a parallel implementation. *Submitted to Future Generation Computer Systems*.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

[16] H. Gonzalez, A. Y. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, W. Shen, and J. Goldberg-Kidon. Google fusion tables: web-centered data management and collaboration. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1061–1066. ACM, 2010.

[17] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7):1645–1660, 2013.

[18] A. Halevy. Best-effort modeling of structured data on the web. In *Conceptual Modeling–ER 2011*, pages 32–32. Springer, 2011.

[19] A. Halevy, A. Rajaraman, and J. Ordille. Data integration: the teenage years. In *Proceedings of the 32nd international conference on Very large data bases*, pages 9–16. VLDB Endowment, 2006.

[20] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

[21] B. He and K. C.-C. Chang. Statistical schema matching across web query interfaces. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 217–228. ACM, 2003.

[22] A. Labrinidis and H. Jagadish. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12):2032–2033, 2012.

[23] D. Laney. 3-d data management: Controlling data volume. *Velocity and Variety, META Group Original Research Note*, 2001.

[24] J. Peral, A. Ferrández, R. Tardío, A. Maté, E. de Gregorio, D. Gil, and J. Trujillo. Energy consumption prediction by using an integrated multidimensional modeling approach and data mining techniques with big data. *Submitted to Future Generation Computer Systems*.

[25] O. Sangupamba Mwilu, N. Prat, and I. Comyn-Wattiau. Business intelligence and big data in the cloud: Opportunities for design-science researchers. *Submitted to Future Generation Computer Systems*.

[26] P. Venetis, A. Halevy, J. Madhavan, M. Paşca, W. Shen, F. Wu, G. Miao, and C. Wu. Recovering semantics of tables on the web. *Proceedings of the VLDB Endowment*, 4(9):528–538, 2011.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

David Gil is an associated professor at the Department of Computing Technology and Data Processing at the University of Alicante, Spain. David received a Ph.D. in Computer Science from the University of Alicante (Spain) in 2008. His research interests include Applications of Artificial Intelligence, data mining, data warehouses, multidimensional databases, OLAP, design with UML, and MDA. He has published papers in high quality international conferences such as IJCNN, SAC, HEALTHINF, DCAI, SCAI, SAIS, etc. He has also published papers in highly cited international journals such as Expert Systems With Applications, Applied Soft Computing. Dr. Gil has served as a Program Committee member of several conferences and workshops such as DAWAK, ARES and CAiSE. He is a reviewer of several journals such as Neurocomputing, Expert Systems and Soft Computing. He is also involved in the organization of several international workshops (MoDIC'12, MoBiD'13-14).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Dr. Il-Yeol Song is professor in the College of Computing and Informatics of Drexel University and Director of Ph.D. Program in Information Studies in his college. He served as Deputy Director of NSF-sponsored research center on Visual & Decision Informatics (CVDI) between 2012-2014. He is also an affiliated professor of Computer Science Department of KAIST, Korea. He is an ACM Distinguished Scientist and an ER Fellow. He is the recipient of 2015 Peter P. Chen Award in Conceptual Modeling. His research interests include conceptual modeling, data warehousing & OLAP, big data management & analytics, CRM, object-oriented analysis & design, healthcare informatics, and smart health. Dr. Song published over 190 peer-reviewed papers and co-edited 22 proceedings. He is a co-Editor-in-Chief of Journal of Computing Science and Engineering (JCSE) and is in an editorial board member of DKE, JDM, IJEBR, and JDFSL. He won the Best Paper Award in the IEEE CIBCB 2004. He won 14 research awards from competitions of annual Drexel Research Days. He also won four teaching awards from Drexel, including the most prestigious Lindback Distinguished Teaching Award. Dr. Song served as the Steering Committee chair of the ER conference between 2010-2012. He is a steering committee member of ER, DOLAP, BigComp, and ADFSL conferences. He served as a program/general chair of over 20 international conferences/workshops including DOLAP'98-14, CIKM'99, ER'03, FP-UML'06, DaWaK'07-'08, , DESRIST'09, CIKM '09, MoDiC'12, and MoBiD'13-'15.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

*Biographies (Photograph)

Click here to download high resolution image

ACCEPTED MANUSCRIPT

*Biographies (Photograph)

Click here to download high resolution image

- Objectives of the third International Workshop on Modeling and Management of Big Data (MoBiD'14).
- Summary of the selected papers.
- Conceptual modeling in the big data era.
- Expectation in these topics for this and the next editions of this workshop.