Journal of Theoretical and Applied Information Technology
20th June 2016. Vol.88. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A BRIEF ON BIG DATA: LESSONS FROM 10 YEARS OF EXPERIENCE

[1,4]DIMITRIOS XANTHIDIS, [2]PAVLOS NIKOLAIDIS, [3]OURANIA KOUTZAMPASOPOULOU

[1]Dhofar university, Department of Management Information Systems, College of Commerce and Business Administration, Oman

[2]IMAM university, Department of Information Systems,

College of Computer and Information Science, Saudi Arabia

[3]Universiti Malaya, Department of Information Systems, College of Computer Science, Malaysia

[4]CIBER-research.eu UK

E-mail: [1]dxanthidis@du.edu.om, [2]paul@ccis.imamu.edu.sa, [3]xarania@gmail.com

ABSTRACT

In this review study, one of the most important technological trend, i.e. big data, is explained together with some of the main terminology relative to the field emerging from it. Some of the challenges and opportunities that it brings along are described and a list of new research sub-fields is also suggested as well as recommendations for the universities to embrace the new technologies and incorporate them in their curriculum in an effort to properly train the experts of the new big data era.

Keywords: *Big Data, 6 Vs, Sources of Data, Data needs, Challenges, Opportunities, New fields in IT*

1. INTRODUCTION

No change has ever come without doubts as to its feasibility and value. Questioning a new idea and its application is not only something everyone should expect but, moreover, useful and reasonable to do as a way to, quickly, pass the period of hype and reach that of its effective use. This is likewise with the concept of Big Data one of the hottest fields of Information Systems and Technology that have emerged the past decade.

Everyone in the IT/IS field is talking about it and most of the professionals in the business fields seem to know everything about it, even if they really don't. It is, as it seems, the easy answer on what topic should a university or organization select as the main theme of its planned conference or at least it is one of the main topics, of course with a good reason. Literally thousands of papers have been published in top journals of IEEE, ACM, Elsevier, Wiley, Taylor & Francis, Emerald, Springer not to mention the other tens of thousands of papers published in other publications. When one goes through them (rather some of them) the feeling is they form the scene of a very big discussion on what it is and why (not if) it is so important.

It involves, seemingly, everyone in the developed or developing world and it, also seems, to affect all of us one way or another. The business people treat it as being the secret to get the competitive advantage against their competitors. The societies are trying to understand how they can benefit from it and at the same time seek for ways to shield and protect their values against its negative consequences. The universities in the developed countries have already a department dedicated to the field, whether with this name or very similar like Data Science, and those in the developing countries either follow or make plans to follow that trend even if they don't know what the requirements are for such a move or even when they don't have the faculty members with the expertise to cover it in a satisfactory way.

Big data is sanctified by many as being a must when trying to solve complex contemporary Information Systems or Information Technology problems. Finding the proper way to manipulate it has become similar to the quest for the Holy Grail. At the same time it raises so many issues to all types of businesses and the various societies that several demonize it or, at the very least, they

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

tend to undermine it as being an old idea offered in a new wrapping.

Despite being a relatively new field, about 10 years old, it has already achieved to shake the business and ICT community and the developed and/or developing societies worldwide as a large magnitude earthquake threatening to reshape all aspects of their understanding, behavior and activity.

2. AIM AND OBJECTIVES

The aim of this review is to inform about the most important terms and issues closely related to Big Data as they appear to be accepted by the international business and IT/IS community. More specifically its objectives are to:

- Explain, in a way as simple as possible, the main terminology associated with the field,
- Seek the origins of the field, i.e. what were (are) the main reasons that dictated its emergence as a necessity,
- Understand the challenges and the opportunities that appear from its use and application,
- Brief about some of the many new sub-fields formed as relevant, directly or not, to Big Data.

Published papers in top journals in the IT/IS field are used to explore the subject in this review paper and inform the reader of the lessons and experiences gained from almost 10 years of developments in this new field.

3. RELEVANT TERMINOLOGY

It is difficult to find a unanimous verdict of the definition of Big Data. There are, most likely, two reasons for that. One is that most researchers and professionals define it by explaining its main characteristics instead of directly explaining its nature. The second is that the number and the nature of these characteristics is not finally established and decided since new characteristics are added during these past 10 years that the application and research on the field have rapidly grown. Generally, according to Chen *et al.,* "one could define big data and big data analytics as technologies (e.g. database and data mining tools) and techniques (e.g. analytical methods) that a company can employ to analyze large scale, complex data for various applications intended to augment firm performance in various dimension. With that definition, high-tech data storage, management, analysis capability, and visual technologies are all part of big data analytics" [1].

Hence, Big Data is often defined as data that has the following characteristics usually referred to as the 6 V's (by the way, just a couple of years ago they were just 3) [2]:

- Volume: mind-bungling magnitude of data in the exascale, usually reported in terabytes and petabytes or even more stored per day (!) that all need to be processed fast if not instantly,
- Variety: excessive heterogeneity in a dataset, seldom structured (i.e. in a tabular form) although preferred, usually either semi- structured or completely unstructured,
- Velocity: very high rate at which data are generated and difficult to manage as to the speed that it should be analyzed and acted upon in order to be of some serious value,
- Veracity: difficult to control unreliability, according to IBM, as a big part of this data comes from sources, like the Social Media, which are uncertain in nature and/or characterized by human subjective judgment,
- Variability: intense variability (some, like SAS, call it complexity) of the data flow rates, with frequent peaks and troughs coming from an unprecedented and, often, unpredicted number of sources all of which needs to be cleansed and, possibly, transformed in a format useful for further analysis to take place,
- Value: according to Oracle, data received in the original form usually has a low value (if at all), which, though, increases as its volume increases.

Each of the aforementioned characteristics of big data could be enough to produce, and indeed does, new research fields and offer solutions and opportunities to various disciplines across a variety of scientific areas including, but not limited to, psychology, sociology, anthropology, computer science, mathematics, physics, economics, and marketing [3].

The main issue is that it causes the need to use far more advanced techniques and technologies to address data analysis needs that come from the past. The following is a short list of such needs as explain by Seltzer [4] but is, by far, not exhaustive at all:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Data Warehousing: since every customer transaction is recorded in detail the businesses have the opportunity to analyze their demographics and behavior and improve their marketing strategies,
- Directory services: "the prevalence of multivalued attributes makes a relational representation quite inefficient",
- Web search: effective indexing that leads to advanced and effective Internet search engines is at the core of the success of companies like Google serving the need of their customer to find in real-time what they are looking for,
- Mobile device caching: making a small size, limited RAM, relatively small processing power CPU able to process the data required through the mobile devices is, by itself, a challenge,
- XML management: online transactions are increasingly being conducted by exchanging XML-encoded documents; since the size and volume of this documents increases rapidly, perhaps exponentially, the processes followed must be also dramatically improved,
- Stream processing: data is not produced at the same source where it is also processed and the old technologies, e.g. relational databases, are not well-equipped in handling all these data with the characteristics listed previously; new technologies are needed for that purpose.

There is a flood of new applications produced (or planned) to address all these emerging needs. They affect different fields in different ways like the following list explains [1]:

- eCommerce and Market intelligence: "recommender systems, social media monitoring and analysis, crowd-sourcing systems, social and virtual games",
- eGovernment and politics 2.0: "ubiquitous government services, equal access and public services, citizen engagement and participation, political campaign and e-polling",
- Science & technology: new ways of getting actual data, new systems for analysis and interpretation, improved procedures to discover knowledge,
- Smart health and wellbeing: "human and plant genomics, healthcare decision support, patient community analysis,
- Security and public safety: "crime analysis, computational criminology, terrorism, open- source intelligence, cyber security.

It goes without saying that such dramatic paradigm shifts in the various disciplines could not come without challenges and disputes on various particular aspects of its applications but, also, offer new opportunities with the potential of great benefits to whole societies worldwide.

4. CHALLENGES ASSOCIATED WITH BIG DATA

There are several challenges that are, one way or another, associated with the advent of the Big Data and its increasing importance and use. One of them is the technology that is needed to effectively process these data. This is one of the major issues related to the processing of Big Data, i.e. whether traditional, well-established and certain legacy relational database management systems can offer, perhaps with some fine-tuning for the occasion and based on Big Data needs, the solutions that are appropriate and required or, instead, if new technologies are more up to this task. In other words, the question is whether SQL-based, RDBMSs are effective in this cases or the new Map-Reduce, proposed and rapidly accepted by all- the-more professionals, and its new architecture is "the new blood" in the developments of the new technology [5].

There are strong arguments in favor and against each side with renowned scholars and professionals worldwide to their support. Generally the argument in favor of the established RDBMS systems has it that these systems have the advantage as to performance and can easily, with only some adoptation of their configuration, manage to effectively process all these data. On the contrary, the argument continues, Map-Reduce related technology have not proved yet to be more effective as the benchmarking has shown in several cases and, indeed, they are not new technologies but only old ideas in new wrappings [6, 7].

The counter-argument is that the traditional RDBMSs are very difficult to configure for many of the Big Data processing needs, they carry big overheads and the costs are overwelming when compared with the actual needs of the companies. On the other hand, Map-Reduce technologies are much more effective, light and easy to install and implement, less costly and require far less technical expertise from the programmers/analysts [5].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The second challenge to be addressed by many businesses is not if there is value in Big Data but what this value is for each of the companies interested in taking advantage of it. Put it simply, the question is not if there is value in Big Data, which is already accepted as a fact by the international business and academic community, but if it is worth for a particular company, with interest in getting involved, to address the several challenges associated with it. Not all the companies have the size and financial means to compete in this arena since they will have to tackle 10 serious technical challenges explained in detail in [8, 9, 10] by Geist and the U.S. Department of Energy and listed below:

- Energy-efficient circuit, power, and cooling technologies are necessary to run big-data infrastructure, and their energy consumption is measured in hundreds of megawatts,
- High-performance interconnect technologies in the exascale big-data computing paradigm cost more to move data around the network as needed than it costs to perform even complex calculations,
- Advanced memory technologies are necessary to improve capacity in an effort to minimize data movement and energy consumption,
- Scalable system software that is power and failure aware is of critical importance as it is generally accepted that the more a business approaches the exascale level the more frequent the failures are and the bigger will be the need for systemic resilience,
- Data management software that can handle the volume, velocity, and diversity of data will play, no doubt, an important role in (trying to) avoid I/O bottlenecks that will limit the system utility and applicability,
- Programming models to express massive parallelism, data locality, and resilience are a must since in exascale level computing systems will have billion-way parallelism and frequent faults,
- Reformulation of problems and refactoring solution algorithms may be necessary so as to save many thousands of person-years of work on big-data related tasks in the company,
- Ensuring correctness in the face of faults, reproducibility, and algorithm verification will be the only way to face frequent and permanent faults,
- Mathematical optimization and uncertainty quantification for discovery, design, and decision are the core of the big-data problems since uncertainty is one of its key characteristics,
- Software engineering and supporting structures are valuable in enabling the researchers increase their productivity aiming to address the all-the-more growing needs for computation related solutions.

Another challenge, that many professionals and academics, either consiously or not, seem to often neglect is the issue of privacy and anonymity that is closely associated with the advent of big data and, if not properly addressed, will cause difficulties for the societies to cope with. As Skopek suggests *"... under the condition of privacy, we have knowledge of a person's identity, but not of an associated personal fact, whereas under the condition of anonymity, we have knowledge of a personal fact, but not of the associated person's identity. In this sense, privacy and anonymity are flip sides of each other. And for this reason, they can often function in opposite ways: whereas privacy often hides facts about someone whose identity is known by removing information and other goods associated with the person from public circulation, anonymity often hides the identity of someone about whom facts are known for the purpose of putting such goods into public circulation*" [11, 12].

In light of the aforementioned statement and given the huge volumes of data, some of which personal and private, it is necessary to address a key issue. Quite frequently the policies related to informed consent from the user/consumer on the one hand and the company's intense interest in addressing personal customer needs and particular behaviors on the other, have led to a clash of two of the most important contemporary concepts, namely that of personal privacy and the one on business marketing strategy. This clash has, in turn, led to what is called the "transparency paradox" that suggests simplicity and fidelity cannot both be achieved at the same time. And, guess what, when comparing the need to ensure an individual's personal privacy and anonymity with the needs of the businesses (in light of the big data) and the societies as a whole, it looks so far, the privacy and anonymity "loses the trade-off with big data" [13].

This is not an issue to take lightly. If the societies and the governments worldwide don't take it seriously and neglect to find ways to cope with then

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Orwellian dark scenarios might just realize in the, not so far, future.

5. OPPORTUNITIES AND NEW RESEARCH FIELDS

Despite the possible problems that any business, government, or other entity is, most likely, to face when involved with big data the verdict is all the more unanimous: there are benefits for those that will be able to cope with these obstacles and utilize big data. However, it is not trivial to identify these benefits since they are expected to differ from sector to sector [14].

The primary objective behind the use of big data in industrial applications is to achieve a fault-free and cost efficient running of the process, while realizing the desired performance levels, especially with respect to quality. McKinsey suggests that manufacturers could make up to a 50% decrease in product development and assembly costs, and up to a 7% reduction in working capital through the use of big data [15].

In sales, big data can add value and shed light in areas "*such as product and market development, operational efficiency, market demand predictions, decision making, and customer experience and loyalty*" according to the Information Systems Audit and Control Association [16].

Generally, a study carried out by IBM [17] revealed the following five main objectives that the professionals but also consumers would expect to benefit from big data:

- Customer-centric outcomes: 49%
- Operational optimization: 18%
- Risk/financial management 15%
- New business model 14%
- Employee collaboration 4%

The concept looks so promising that even the governments of highly developed countries like the U.S., Japan, China, Germany and others have engaged in funding large projects related anyhow with it [18]. In the case of Germany, in particular, the authorities and the decision makers are so confident of the critical role that Big Data will play in giving a competitive advantage in the local industry against that in the rest of the world that they decided to implement the strategic initiative named "Industrie 4.0" that will lead them, first before any other industry in the world, to the so- called "fourth industrial revolution" [19].

There is a large number of fields and applications associated with Big Data but, generally, they are divided into 4 different groups based, mainly, on the actual phase of the big data pipeline, namely, data generation, data acquisition, data storage, and data analytics [20]. Even in each of the particular groups, though, there are numerous possibilites for research and specialization. For instance, in the case of data analytics the category can be further divided into text analytics, video analytics, audio analytics, social media analytics, predictive analytics and others (possibly). Likewise in the case of data generation, since there is a huge variety of sources from which data may originate including social media and online transactions [2].

The above realities point to another fact that, sooner or later, societies will, also, have to face. Since big data is reshaping all aspects of everyday life there will be very soon a great need for specialists in the various sub-fields. Indeed, a 2012 study predicted that by 2018, the U.S. will need close to 200,000 people with deep data analytics specialization and an additional 1.5 million data-savvy managers with the know-how to analyze big data to make effective decisions [21]. This need was expressed quite clearly by Hal Varian, Chief Economist at Google and Emeritus Professor at the University of California, Berkeley: "… so what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis. So my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on" [22].

Chen *et al.* [1] provided a list of emerging research areas implying the specialists that will be required by the industry (various sectors) and the academia not just in the future but even starting today. Some of the most trendy (a quick look at the conference themes of the past few years easily points to them) are:

- Web mining and network mining,
- MapReduce/Hadoop,
- Column-based DBMS and parallel DBMS,
- Information extraction,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Opinion mining,
- Sentiment/affect analysis,
- Visualization,
- Multimedia and mobile IR,
- Cloud services and cloud computing,
- Social Media analytics,
- Web visualization,
- Social marketing,
- Web privacy/security,
- Virtual communities,
- Trust and reputation,
- Mobile web services,
- Personalization and behavioral modeling.

Add to the above the areas of virtualization, grid computing, green computing, quantum computing and the reader will get a pretty good idea of where the IT/IS technology will move towards in the next years.

6. DISCUSSION - CONCLUSIONS

Nearly two centuries ago, the English chemist Humphrey Davy wrote: "Nothing tends so much to the advancement of knowledge as the application of a new instrument. The native intellectual powers of men in different times are not so much the causes of the different success of their labors, as the peculiar nature of the means and artificial resources in their possession" [8].

Big Data came to stay for good. It involves, it issafe to say, all of us one way or another and it affects all aspects of everyday life. Those who believe it is an old concept in a new wrapping are, quite likely, correct. However, this new wrapping is absolutely transparent allowing us to see two completely different realities. One, look in the past of technology, and Information and Communication Technology in particular, and see our many, as it seems, weaknesses and imaturities. Two, have a glimpse of the future and allow us the time to proceed with policies, planning and actions that will lead us to this new era that has emerged and, though still uncharted area, prepare the tools to walk through it in a way as safe as possible for our societies.

The ICT is, in many ways that are out of the scope of this paper, responsible for many of the digital divides worldwide but, when societies are properly educated, it is at the same time the main instrument to bridge these digital divides. Education of the societies, actually higher education, is the vehicle towards preparing the peoples in the new era that is upon us. Universities have a pivotal role to play in this. It is only required for their authorities to be open minded and quickly realize the new needs.

REFERENCE

[1] H. Chen, R. H. L. Chiang, V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact", *MIS Quarterly,* 36(4), pp. 1165-1188, 2012.

[2] A. Gandomi, M. Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics", *International Journal of Information Management,* Vol. 35, pp. 137-144, December 2014, DOI: 10.1016/j.ijinfomgt.2014.10.007.

[3] W. He, S. Zha, L. Li, "Social M edia Competitive Analysis and Text Mining: A Case Study in the Pizza Industry", International Journal of Information *Management,* Vol. 33(3), pp. 464-472, 2013. [4] M. Seltzer, "Beyond Relational Databases", *Communications of the ACM,* Vol. 51, No. 7, pp. 52-58, July 2008, DOI:10.1145/1364782.1364797.

[5] M. Stonebraker, D. Abadi, D. J. Dewitt, S. Madden, E. Paulson, A. Pavlo, A. Rasin, "MapReduce and Parallel DBMSs: Friends or Foes?", *Communications of the ACM,* Vol. 53, No. 1, January 2010, pp. 64-71, DOI: 10.1145/1629175.1629197.

[6] M. Stonebraker, "Big Data is 'Buzzword du Jour;' CS Academics 'Have the Best Job', *Communications of the ACM,* Vol. 56, No. 9, September 2013, pp. 10-11, DOI: 10.1145/2500468.2500471.

[7] A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. Madden, M. Stonebraker, "A Comparison of approaches to Large-Scale Data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Analysis", Proceedings of the 2009 ACM SIGMOD International Conference (Providence, RI, June 29 – July 2), ACM Press, New York, 2009, Available at: http://database.cs.brown.edu/projects/mapredu ce-vs-dbms/.

[8] A. Geist, R. Lucas, "Major Compute Science Challenges at Exascale", *International Journal of High Performance Applications,* Vol. 23, No 4, November 2009, pp. 427-436.

[9] R. Lucas, J. Ang, K. Bergman, S. Borkar, *et al.,* "Top Ten Exascale Research Challenges", Office of Science, U.S. Department of Energy, Washington, D.C., Available at: http://science.energy.gov/ ~media/ ascr/ ascac/ pdf/ meetings/20140210/Top10reportFEB14.pdf.

[10] U.S. Department of Energy, "The Opportunities and Challenges of Exascale Computing", Office of Science, Washington, D.C. 2010, Available at: http://science.energy.gov/~media/ ascr/ ascac/ pdf/ reports/ Exascale_subcommittee_report.pdf.

[11] J. M. Skopek, "Anonymity, the Production of Goods, and Institutional Design", *Fordham Law Review,* 82, 4, 2014, pp. 1851-2809, Available at: http://ir.lawnet.fordham.edu/ flr/ vol82/ iss4/4/.

[12] J. P. Daries, J. Reich, J. Waldo, E. M. Young, J. Whittinghill, A. D. Ho, D. T. Seaton, I. Chuang, "Privacy, Anonymity and Big Data in the Social Sciences", *Communications of the ACM,* Vol. 57, No. 9, September 2014, pp. 56-63, DOI: 10.1145/2643132.

[13] S. Barocas, H. Nissenbaum, "Big Data's End Run Around Procedural Privacy Protections", *Communications of the ACM,* Vol. 57, No. 11, pp. 31-33, DOI: 10.1145/2668897.

[14] S. Yin, O. Kaynak, "Big Data for Modern Industry: Challenges and Trends", *Proceedings of the IEEE,* Vol. 103, No. 2, pp. 143-146, February 2015, DOI: 10.1109/JPROC.2015.2388958.

[15] McKinsey, "Big Data: The Next Frontier for Innovation, Competition, Productivity", McKinsey Global Institute Report, 2011.

[16] Information Systems Audit and Control Association (ISACA), "Big Data: Impacts and Benefits", March 2013, White Paper. [17] IBM and Said Business School, "Analytics: The Real-World Use of Big Data: How Innovative Enterprises Extract Value from Uncertain Data", IBM Institute for Business Value and Said Business School Executive Report, October 2012.

[18] D. A. Reed, J. Dongarra, "Exascale Computing and Big Data", *Communications of the ACM,* Vol. 58, No. 7, pp. 56-68, July 2015, DOI: 10.1145/2699414.

[19] H. Kagermann, W. Wahlster, "Securing the Future of German Manufacturing Industry: Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0", Working Group, Acatech – National Academy of Science and Engineering, Germany, 2013, Final Report of the Industrie 4.0.

[20] J. Gantz, D. Remsel, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East", *Proceedings of IDC iView,* IDC Anal. Future, 2012.

[21] J. Manyika, M. Chui, B. Brown, J. Bughin, R.

Dobbs, C. Roxburgh, A. H. Byers, "Big Data: The Next Frontier for Innovation, Competition, and Productivity", McKinsey Global Institute, Available at: http://www.mckinsey.com/ insights/ mgi/ research/ technology_and innovation/ big_data_the_next_frontier_for_innovation.

[22] H. Varian, "Hal Varian Answers your Questions", Freakonomics, 25/8/2008, Available at: http://www.freakonomics.com/2008/ 02/ 25/ hal-varian-answers-your- questions.