European Journal of Business and Social Sciences, Vol. 3, No. 5 , pp 59-66, **August 2014.** **P.P. 59 - 66**
URL: *http://www.ejbss.com/recent.aspx*
ISSN: 2235 -767X

EXTRACTING KNOWLEDGE FROM DATA: FROM BUSINESS INTELLIGENCE TO BIG DATA ANALYTICS

Giuseppe Pirlo,
Università degli Studi di Bari, via Orabona 4, 70125

Donato Impedovo,
Politecnico di Bari, via Orabona 4, 70125

Lorenzo Scianatico,
Università degli Studi di Bari, via Orabona 4, 70125

Annalisa Vacca
Università degli Studi di Bari, via Orabona 4, 70125

ABSTRACT

At current time, a revolution has started. A revolution called Big Data. This revolution is not limited to volume of data: this revolution requires us to change everything about technology those we daily use. It is necessary to change the architecture of modern data centres, the programming technologies available, and the way we see (and use) sensors and other instruments, and finally, requires changing the application of Computer Science and Engineering to these data.

The aim of this work is to how this change is big and how important it is. Moreover, perspectives, challenges and future work directions will be examined.

Key words: knowledge, analytics, big data, business intelligence

1. Introduction

Let us introduce the main protagonist of one of many modern era revolutions: Big Data. This is the name that currently rules a good part of scientific research. The greatest effort is to reach any possible advantage from collecting this so called Big Data: this means, data harvesting, data management, data manipulation and data interpretation. In few year data analysts have seen the amount of data grown exponentially. Few years ago it was sufficient using some simple and basic statistic to evaluate data, like mean or variance or other indexes. But probably no one was able to imagine how data could increase their dimensions, and most of all, their importance. Modern technology brought us the greatest data source of ever: these sensors are continuously measuring every type of information at the highest rate is reachable, and continuing sending them to a processor. A great part of these data is stored in various way, like databases or some kind of files, well-structured and unstructured. And then, they became Big Data. What are the most important aspects of big data?

2. Definition of Big Data

It is a well-known fact the amount of data generated on Earth is growing faster and faster. It has been calculated that in 2010 the total amount of data was equal to one Zettabyte [Zaslavsky et al. 2013] and in 2011 was equal to 1.8 Zettabyte. Some estimations says that in 2020 the total amount will be equal to 35 ZB, but someone says that this estimation may be too conservative. Also, the same report made in the Australian agency CSIRO shows how there is a gap between the data that an organization can process and the data available to the same organization. This is due to the lack of data analysis technology in any organization that does not consider how many scientific field are now became data driven disciplines.

The same work shows an outline of typical Big Data characteristics. These are known as the three V: Variety, Volume and Velocity.

- Variety refers to the heterogeneous type of data: these data, as mentioned before, are stored in various ways, like relational DBMS, text file (both structured and unstructured), new generations DB, like MongoDB or Neo4J, graph oriented databases and so on. Moreover, these data are various by typology: they are sensed by a great variety of sensors and this means that are very different values in same datasets.
- Volume refers to the amount of data. Big Data size goes from Terabytes to Petabytes and sometimes to ZettaBytes.
- Velocity refers to processing modality: it is time to dismiss batch processing to adopt stream processing. Rapidity of results is essential to use correctly Big Data.

It is immediate to understand that classical computer and simple architectures are not sufficient to enforce data analysis on datasets that became more and more complex. Business Intelligence has to evolve and to move from simple descriptive statistics to inferential statistics, and it is necessary adopt new techniques to describe all phenomena “inside” data.

The foundation of data analysis resides in data warehouses: with technology of first RDBMS and first generation of data mining techniques has started a new evolution. Then, the advent of World Wide Web gave everyone a first, unique ability to access easily to great datasets among the world. The introduction of search engines, like Google, gave access to data logs everywhere, and opened all various databases to the Web connection. At last, the introduction of new technologies like tablets, smartphone, and the improvement of sensors reached a new kind of importance in data production, which has opened new ways of scientific research and new development directions. Currently, there is no one that is not involved in this great change, who has great names interested, like Google, eBay and Amazon. [Chen et al., 2012]

The greatest part of data is analyzed in a scientific context: meteorology, genomics, physics and environmental science are primary stakeholders of this technology, who has evolved in these last years. Let us see how.

3. History of Big Data

As shown by [Aronova et al., 2010, pp 183-224], two events after World War II started the idea of Big Science: the International Geophysical Year (1957) and the International Biological Program (1964-1974). Together with the Manhattan Project and other space programs, these events outlined a very long-term program that today lead to Big Data. However, it is necessary to remark that the first two event did not reach results in short time. Remember that technology was not sufficient to provide what is necessary to analyze data. However, IBP proposed a new model of Big Science that became the Long-Term Ecological Program, which inspired various scientific programs all over the world. Starting from Geophysical science, the idea passed into Biological sciences and passed in Social studies. In this way, a first framework for Big Science evolution was created. During years, instruments were changed. In fact computer processing capacities were improved, and simultaneously computer science technique became more and more refined an efficient. Main differences between old Business Intelligence and modern Big Data Analytics can be traced as it follows [op. cit.]:

- The first generation of Business Intelligent was strongly RDBMS-oriented and organized as well-structured content. Typical use of these systems include ad-hoc query, search-based BI, reporting, OLAP, predictive modeling and data mining.
- Second generation of Business intelligence was changed by the Web advent. Principal innovation was the use of less structured content and the adoption of web mining, social media and network analysis, spatial and temporal analysis, several techniques of information retrieval and extraction, semantic information processing.
- The new third generation BI is more sensor oriented. It focuses on person-centered, location-aware analysis, and is more orients as Mobile BI.

The mayor improvement for BI comes from new technologies like smartphone and tablet. These devices are equipped with many sensors, which produces constantly new data. In addition, new development platforms like Arduino and Raspberry Pi are launching a new revolution known as “makers’ movement”. Currently, many people develops devices like 3d-printers, home automation, drones and robots and so on, with these new low-cost platforms.

Data fusion techniques are becoming more and more important. Each sensors produces different types of data, which is typically stored in various way. These data can be stored in a well-structured way, like traditional DBMS (notice that even Android has an internal DBMS, Sqlite), or XML-based files, or unstructured way, like simple text-file. Therefore, it is important to re-organize data.

In this new perspective of person-based analysis another technological paradigms are emerging [Bughin et al., 2010, pp. 75-86]: together with the Internet-of-things, now it is possible to imagine everything as a service.

4. Applications of Big Data

Although Big Data is a new discipline, there are a lot of application field, which are suitable for them. There are various applications, challenges and success examples [Bizer et al., 2012, pp 56-60].

In 2011, four researchers showed four challenges of Big Data:

- The first one is about data integration. Integrating real, messy, schema-less data must be enforced by multi-technology and multi-disciplinary approaches.
- The second one is about RDF processing: a concrete task, which touches all challenges around data integration, large-scale RDF processing, and data quality assessment that arise in the context of the Web of Data, is to understand semantically all data harvested through various sources.

- The third one is about a technology that can rip all possible valuable data from Linked Open Data, which is called by its own creator "LOD ripper".
- The fourth one is to let data integration to drive DBMS technology, natively incorporating semantic information of data.

As said before, there are many application fields of Big Data. Main applications are [Villars et al., 2011]:

- Media/entertainment: The media/entertainment industry moved to digital recording, production, and delivery in the past five years and is now collecting large amounts of rich content and user viewing behaviors.
- Healthcare: The healthcare industry is quickly moving to electronic medical records and images, which it wants to use for short-term public health monitoring and long-term epidemiological research programs. At current time, US federal government is transforming their healthcare systems, opening to new techniques of storage and new analysis methodologies [Kayyali et al., 2013].
- Life sciences: Low-cost gene sequencing (<$1,000) can generate tens of terabytes of information that must be analyzed to look for genetic variations and potential treatment effectiveness.
- Video surveillance: Video surveillance is still transitioning from CCTV to IPTV cameras and recording systems that organizations want to analyze for behavioral patterns (security and service enhancement).
- Transportation, logistics, retail, utilities, and telecommunications: Sensor data is being generated at an accelerating rate from fleet GPS transceivers, RFID tag readers, smart meters, and cell phones (call data records); that data is used to optimize operations and drive operational business intelligence (BI) to realize immediate business opportunities.
- Many scientific disciplines have become data-driven. For example, a modern telescope is really just a very large digital camera. The proposed Large Synoptic Survey Telescope (LSST) will scan the sky from a mountaintop in Chile, recording 30 trillion bytes of image data every day. Astronomers will apply massive computing power to this data to probe the origins of our universe. The Large Hadron Collider (LHC), will generate 60 terabytes of data per day – 15 petabytes (15 million gigabytes) annually. The Italian satellite Cosmo-Skymed produces an enormous amount of information (about 1800 images per day) that requires great storage and great processing to obtain added-value products[Bianchessi e Righini, 2008, pp 535-544; Coletta et al, 2008, 5-13]. Generally, all remote sensing application are distinct by great effort on data processing [Benz et al., 2004, pp 239-258]. Similar eScience projects are proposed or underway in a wide variety of other disciplines, from biology to environmental science to oceanography and meteorology. These projects generate such enormous data sets that automated analysis is required. Additionally, it becomes impractical to replicate copies at the sites of individual research groups, so investigators pool their resources to construct a large data center that can run the analysis programs for all of the affiliated scientists.
- Modern medicine collects huge amounts of information about patients through imaging technology (CAT scans, MRI), genetic analysis (DNA microarrays), and other forms of diagnostic equipment. By applying data mining to data sets for large numbers of patients, medical researchers are gaining fundamental insights into the genetic and environmental causes of diseases, and creating more effective means of diagnosis. It is also known that biocuration needs urgently structures, support and recognition[Howe et al. 2008, pp 47-50].
- Our intelligence agencies are being overwhelmed by the vast amounts of data being collected through satellite imagery, signal intercepts, and even from publicly available sources such as the Internet and news media.
- The collection of all documents on the World Wide Web (several hundred trillion bytes of text) is proving to be a corpus that can be mined and processed in many different ways. For example, language translation programs can be guided by statistical language models generated by analyzing billions of documents in the source and target languages, as well as multilingual documents, such as the minutes of the United Nations.

5. Software and skills

Even if Big Data is a young field in research, some Big Data software is already available for any professional. Probably the most well-known software is Apache Hadoop, inspired by Google Analytics, but also there are some alternatives. A first interesting alternative is Mapreduce, as shown by Chen, Alspaugh and Katz [Chen et al, 2012, pp 1802-1813].

Within the past few years, organizations in diverse industries have adopted MapReduce-based systems for large-scale data processing, like Facebook and Cloudera. Along with these new users, important new workloads have emerged which feature many small, short, and increasingly interactive jobs in addition to the large, long-running batch jobs for which MapReduce was originally designed. As interactive, large-scale query processing is a strength of the RDBMS community, it is important that lessons from that field be carried over and applied where possible in this new domain. However, these new workloads have not yet been described in the literature. Their key contribution is a characterization of new MapReduce workloads which are driven in part by interactive analysis, and which make heavy use of querylike programming frameworks on top of MapReduce. These workloads display diverse behaviors which invalidate prior assumptions about MapReduce such as uniform data access, regular diurnal patterns, and prevalence of large jobs.

At UC Irvine University [Borkar et al., 2012] is currently developed another Big Data system, known as ASTERIX. The focus of this project is to develop a well-structured architecture for Big Data problems, and their work shows their approach to the Big Data problem.

Starfish [doc19] is a project developed in Duke University. The main focus of the Starfish project is to develop a self-tuning Data Analytics system. Moreover, in this project has been shown new challenges in Big Data development.

This is not a complete list. There are lots of tutorials on the network, like the one shown by Agrawal. [Agrawal et al., 2011]

Software is not the only important requirement in Big Data development. Recently, in the list of professional of the future there is a new entry, sometimes called “Data Scientist”, or “Data Engineer” or “Analytics Engineer”. For these new professional it is important not only to have good software instruments, but even skills. In the new data-driven world, the theoretical basis of these new professionals is made by all data disciplines: statistics, probability, numerical analysis, econometrics, machine learning, data mining and information retrieval and so on. In addition, it is important to have a good capability with databases, like the relational ones and the various NOsql systems. Someone, like Zaslavsky, has proposed to create a BS or MS degree in Data Analytics, or alternatively, a good background in Computer Science and a postgraduate course in Data Science [op. cit.]. However, the most important feature is to create the so-called MAD abilities in analytics systems [Herodotou et al,. 2011; Cohen et al., 2009, 1481-1492]: MAD stand for:

- Magnetic: a magnetic system attracts all sources of data irrespective of issues like possible presence of outliers, unknown schema or lack of structure, and missing values that keep many useful data sources out of conventional data warehouses;
- Agility: an agile system adapts in sync with rapid data evolution;
- Depth: A deep system supports analytics needs that go far beyond conventional rollups and drilldowns to complex statistical and machine-learning analysis.

The same creator of Starfish also defined the MADDER abilities. The first three letters have the same meaning of the MAD acronyms, while the last three stand for:

- Data-lifecycle-awareness: a data-lifecycle-aware system goes beyond query execution to optimize the movement, storage, and processing of big data during its entire lifecycle;
- Elasticity: An elastic system adjusts its resource usage and operational costs to the workload and user requirements;
- Robustness: A robust system continues to provide service, possibly with graceful degradation, in the face of undesired events like hardware failures, software bugs, and data corruption.

Starfish has been develop to have all of MADDER abilities.

6. Open problems

As all newest technologies, even Big Data has his own problems. In several researches were traced all open problems of current instruments.

Nowadays, various graphics board producer has proposed a new processing paradigm known as GPGPU. This paradigm uses video cards not only for graphics computing, but also for general-purpose computing. Maybe the most famous toolkit is Nvidia CUDA Toolkit, which supports many programming languages, like C, C++, Java, Python and Matlab. Even Intel and AMD have developed an equivalent toolkit. Now, the question is the sequent: is this architecture sufficient for Big Data?

Trelles et al. have investigated this problem. Their work [Trelles et al,. p 224] shows that GPU performances rapidly decreases when large volumes of data are passed. Moreover, GPU are vector processors that are highly suitable only for a subset of computational problems. New trends like Oracle Exadata and IBM Netezza provides better solutions. These platforms have CPUs on the storage itself and in some cases uses an integration of photonics and electronics for improving speed. In particular, it is shown in their work that bioinformatics needs an unprecedented fast storage and calculation speed, maybe abandoning old pure relational database modeling.

At least, other researchers focused their attention on some methodological issue. Speaking of Hadoop [Cuzzocrea et al., 2011]:

- How to build multidimensional structures on top of the HDFS?
- How to directly integrate multidimensional data sources into the Hadoop Lifecycle?
- How to model and design multidimensional extensions of HiveQL?
- How to design complex Analytics over Hadoop-integrated multidimensional data?
- How to deal with visualization issues arising from Big Multidimensional Data Analytics?

These are only example questions, but is very important to give an answer as soon as possible: entire world in this moment is continuing to produce data as fast as he can.

At least, as Madden says, all these issues can be summarized as a unique task [Madden, 2012]: it is necessary to create a complete data management ecosystem to support data scientists when there are experimenting, providing them with technologies, algorithms, visualization instruments to support them in their work.

7. Conclusion

Although Big Data appears to be carrier of great innovation in science, engineering and technology, it is not a no-risk investment. In fact, there is one great risk for Data scientists. The risk is to confront with their work in this way [Lohr, 2012, p. 11]: I know the facts, I have the data to prove the facts I know, so let's find them. Any professional that is confronting with this new millennium challenge has to remember that he is always biased. A good practice to avoid this risk is to not permit Data scientist to work alone. Data scientist should work at least in number of two and should validate their work each other, and it would be better to have another scientist more experienced than the others with the responsibility of final revisioning and validation. This figure can be perfectly covered by the Chief Information Officers [op. cit].

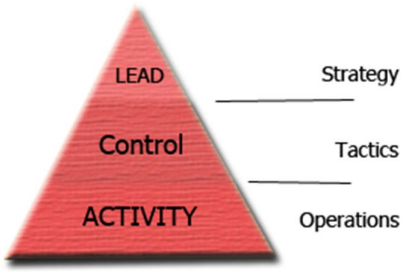

Figure 1: Anthony's Pyramid

The New Information Society will be strongly led by CIOs decisions, which most important effort is making IT more valued asset to the organization. The Big Data opportunity will bring definitively Computer Science from the operational level to the strategic level of the well-known Anthony's Pyramid (Figure 1), considering their importance for the future, becoming strategically partners for future business. In this, CIOs will have a prominent role and will get them more and more involved in business, so they must be prepared to take any advantage is possible.

BIBLIOGRAPHY

1. Agrawal, Divyakant, Sudipto Das, and Amr El Abbadi. "Big data and cloud computing: current state and future opportunities." Proceedings of the 14th International Conference on Extending Database Technology. ACM, 2011.
2. Aronova, Elena, Karen S. Baker, and Naomi Oreskes. "Big science and big data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) network, 1957–present." (2010): 183-224.
3. Benz, Ursula C., et al. "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information." ISPRS Journal of photogrammetry and remote sensing 58.3 (2004): 239-258.
4. Bianchessi, Nicola, and Giovanni Righini. "Planning and scheduling algorithms for the COSMO-SkyMed constellation." Aerospace Science and Technology12.7 (2008): 535-544.
5. Bizer, Christian, et al. "The meaningful use of big data: four perspectives--four challenges." ACM SIGMOD Record 40.4 (2012): 56-60.
6. Borkar, Vinayak, Michael J. Carey, and Chen Li. "Inside Big Data management: ogres, onions, or parfaits?" Proceedings of the 15th International Conference on Extending Database Technology. ACM, 2012.
7. Bughin, Jacques, Michael Chui, and James Manyika. "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch." McKinsey Quarterly56.1 (2010): 75-86.
8. Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS Quarterly 36.4 (2012).
9. Chen, Yanpei, Sara Alspaugh, and Randy Katz. "Interactive analytical processing in big data systems: A cross-industry study of MapReduce workloads." Proceedings of the VLDB Endowment 5.12 (2012): 1802-1813.
10. Cohen, Jeffrey, et al. "MAD skills: new analysis practices for big data. Proceedings of the VLDB Endowment 2.2 (2009): 1481-1492.
11. Coletta, Alessandro, et al. "Il Programma COSMO-SkyMed: descrizione della missione e del sistema e primi risultati." Rivista italiana di telerilevamento 40.2 (2008): 5-13.
12. Cuzzocrea, Alfredo, Il-Yeol Song, and Karen C. Davis. "Analytics over large-scale multidimensional data: the big data revolution!" Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP. ACM, 2011.
13. Herodotou, Herodotos, et al. "Starfish: A Self-tuning System for Big Data Analytics." CIDR. Vol. 11. 2011.
14. Howe, Doug, et al. "Big data: The future of biocuration." Nature 455.7209 (2008): 47-50.
15. Kayyali, Basel, David Knott, and Steve Van Kuiken. "The big-data revolution in US health care: Accelerating value and innovation." Mc Kinsey & Company (2013).
16. Lohr, Steve. "The age of big data." New York Times 11 (2012).
17. Madden, Sam. "From Databases to Big Data." IEEE Internet Computing 16.3 (2012).
18. Trelles, Oswaldo, et al. "Big data, but are we ready?" Nature Reviews Genetics12.3 (2011): 224-224.
19. Villars, Richard L., Carl W. Olofson, and Matthew Eastwood. "Big data: What it is and why you should care." White Paper, IDC (2011).
20. Zaslavsky, Arkady, Charith Perera, and Dimitrios Georgakopoulos. “Sensing as a service and Big Data.” arXiv preprint arXiv:1301:0159 (2013).