Chapter 14

Research and Education in Data Science: Challenges for the Area of Information Systems

Fabio Silva Lopes, Leandro Augusto da Silva, Vivaldo José Breternitz

Abstract

Data Science (DS) is an interdisciplinary research area that uses concepts based on Big Data Analytics, Programming Languages and Mathematical fundamentals to develop research into insight discovery from datasets. The data-driven decision-making approach has become a challenge for researchers and lecturers in the Information Systems area. This is because the skills and theoretical issues require a heavy course workload and high number of class hours. This work introduces a discussion of the insertion of the DS subject in Information Systems courses, and researchers' efforts to establish goals for professionals and lecturers involved in DS.

14.1. Introduction

The proliferation of computational systems in the industry has caused significant changes in the way data are collected, transmitted and analyzed [Turban et al. 2010]. In addition, there is considerable progress in the management of persistent systems and Data Base Management System (DBMS). The advance started with Peter Chen [Chen 1976] who proposed the E-R diagram to aid the design of databases and defined the requirements of problems, entities, attributes and relationships in a graphical way. Complementarily, DBMS has been consolidated as the main database approach, ensuring the properties such as Atomicity, Consistency, Isolation and Durability, called ACID [Silberschatz et al. 2006].

Relational DBMS was the major resource to store data, facts or transactions, and it became a valuable tool for data analysis using tables, reports, maps or graphs, allowing approaches to manage or to monitor business in most areas of knowledge.

Data generation has been growing exponentially and, which demands collecting, organizing, analyzing and extracting insights from data warehouses systematically, in heterogeneous environments, geographically distributed and, in distinct contexts of applications.

Thus, for a professional to be up-to-date with the stack of technology to collect, process, analyze and data visualization is a humanly impossible task. However, at the same time, it can become as a new job.

In the Brazilian context, Information Systems graduates are looking for these opportunities and, the courses coordination often input new contents as part of traditional syllabuses of computing courses.

Considering the need for an interdisciplinary approach to innovation in this area, as well as the approach of an applied science, positioning computer technology as the driving force for innovation, we have identified a major gap in the information systems curricula. There is a lack of focus to perform special skills to meet Data Analysis effectively. For example, the contents of the database syllabus are presented theoretically and systematically. Instead of this, a practical approach can tackle database but emphasize data analysis as well as relationships with other programs. These aspects may be more or less evident depending on how the lecturer deals with the subject. The discussion of different ways of data collection or manipulation of data for using in data mining is usually defined by the scope of the subject used as a case study.

14.2. Background

In recent decades, data storage technologies have undergone progressive advances that have transformed the data collection concept and as a result the ability to generate digital data exceeds the human capacity of analysis.

Before discussing Big Data, it is important to discuss what kind of data this is. We can expand on this point, starting by defining data (from Latin - Datum), which means details. These details represent events observed, which occurred at a certain time and space. They can be collected, organized, classified and analyzed. In the same way, they can be shown in numeric or alphanumeric form, as a symbol, image, etc. The data can be collected from nature or from details about business processes.

Details that can be created day-by-day, as discussed in Simon (1996). The author used the term "artificial" as opposed to "natural" and addressed the world created by human beings, with laws and contexts connected by our collectivity. At this point, we devote especial attention to the dichotomy between natural and artificial.

From nature, there are infinite details to be collected. For example, a bird’s movement can provide data such as velocity, altitude, direction, fly time etc. In the same line, there are details of a job process such as a start date, stakeholders, incomes, descriptions etc., aggregating a set of data that can be the input for analytical studies that will support decisions about specific contexts.

Knowledge is generated from the relationship between the observed object and the human being. In the database context, the data of a referred object, when manipulated in the right way, can produce value judgment and become information. Information used in a context becomes knowledge.

This set of available data, objects attributes, events and processes will compose a large collection, and provide better knowledge about our reality, our context. However, the data desired is not always collected, and the inverse is also true. We do not always have what we need. With or without available data, discovering patterns hidden inside the data in a chaotic universe is a challenge for Computing Science.

There are currently huge volumes of data; so high that it has been necessary to develop techniques to suppress data that does not interfere with a specific analysis, and build models with only essential data for specific purposes. The term Big Data (BD) is used to define a set of technological tools which allows a new approach to treating and exploiting large sets of data from different sources and formats, for decision making purposes (Breternitz et al. 2015). These authors claim that the volume of data has grown exponentially, which has stimulated improvements in tools to treat them and to exploit them.

Big Data is the term currently used to refer to the volume of data we collect from various sources for further mining and for searching non-trivial patterns in a given context. The search for patterns or the deeper understanding of the data collected is usually called Analytics. This is done with the use of different tools, able to process predictive analytics, data mining, statistics, artificial intelligence etc. Certain authors use expressions such as Advanced, Discovery and Exploratory Analytics; from this point on, we use the expression BDA (Big Data Analytics).

According to Vaisman and Zimányi (2014), Analytics can be defined as the discovery and communication of meaningful patterns in data collections. Such patterns become evident through the application of prescriptive and predictive models derived from statistics and artificial intelligence. Identifying patterns helps us take better decisions. However, this is a probabilistic world, and nothing is 100% certain.

We cannot define Big Data as one tool or one product. It is a process with several iterations and requires different skills to perform in a value chain. Figure 14.1 shows the macro process of a Data Value Chain.

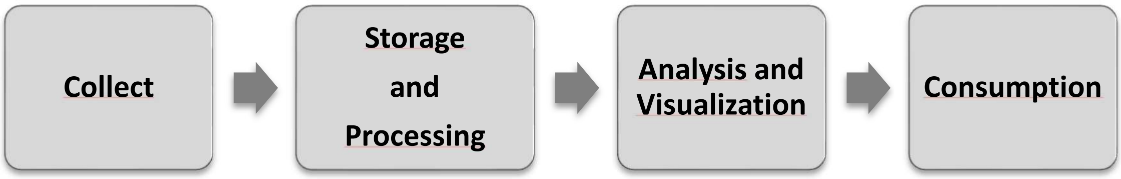

Figure 14.1. Data Value Chain

Building this value chain is not trivial. Big Data analysts are estimated to spend between 50% and 80% of their working time collecting, cleaning and churning data from different sources and formats, in other words, preparing the data to make sophisticated analysis. This has been the focal attention of many startups to build tools based on refined data.

As regards storage and processing, over the years, researchers and computer professionals were observed to lead excellence in the use of DBMSs (Database Management Systems), as well as operating systems with parallel processing capability in high impact performance and scale cluster. This has contributed to consolidating essential transactional systems technologies in several areas such as Economics, Education, Health and Safety.

Data analysis is an interesting topic that generates interest in various contexts. There are different techniques, tools and algorithms for performing analysis and subsequent data visualization. In 1914, William Brinton published a book about graphical methods to show data. We still have the need to find the best way to visualize aggregations and other analytical results.

Finally, data consumption addresses several issues deserving attention. We already have business processes with BDA support in near real-time. It is an advanced stage, but it requires joint efforts from different backgrounds.

Some authors, such as Zikopoulos et al (2012), say that Big Data is characterized by four aspects: Volume, Velocity, Variety and Veracity.

The feature "volume" refers to volumes of data digital form growing exponentially, from not only conventional systems, but also from sources such as Facebook, Tweeter, YouTube, RFID, embedded electronics, cell phones and similar, sensors of various types, etc. To illustrate this feature, McAfee and Brynjolfsson (2012) reported that in 2012, every second, there are more data transiting over the Internet than the total stored in it 20 years ago, and that Walmart alone collected more than 2.5 Petabytes each hour, in order to store transactions made by their clients.

Another important aspect of Big Data is the "velocity" at which data can be captured and processed in nearly real time and this can give a competitive advantage to organizations. This can be exemplified by the experiment conducted by Prof. Alex Pentland's Research Group of MIT Media Lab: the group captured location data of mobiles in order to find out how many people put their cars in the parking lot of the American Macy's group stores on the 2011 Black Friday (which marks the start of the Christmas shopping season in the United States). These sales could be estimated accurately even before they occurred, generating competitive advantages for commercial and marketing areas and third parties, such as investors in stock markets. This case and similar ones are also reported by Clifford (2012).

Regarding "variety", it should be noted that in addition to different sources, the data collected often have different characteristics from those processed by conventional systems, which are not structured and refer to things such as sound, image, movement, temperature, humidity and even variations in the chemical composition of air [Lohr 2012]. Davenport (2014) notes that although the volume of data is what demands the most attention, the lack of structure is the most difficult aspect to work with data storage.

The aspect "veracity" refers to data that cannot be "perfect"; e.g. we must consider how good the data should be to generate useful information and also the cost of making them good.

There have been advances in hardware and software to face demands of performance and availability. It has been a long way to arrive in the Hadoop Platform, which is able to cluster management and parallel processing to leverage BDA. Today, Hadoop is the most widely adopted platform, considering the options available on the market. It is a hybrid model with free and paid associated products, approved by different players in this segment.

New business applications will be using the IoT - Internet of Things. This depends on a universe of data to be collected, processed and stored. At this point, it is important to note that the technologies will be favored by Cloud Computing technologies, in which hardware and software are available to the market as services. The IT business mode is known as XaaS, an acronym, replacing X with P will make "Platform as a Service" or use the I for "Infrastructure as Service."

The issue to be observed is the variety of data that we are dealing with. Different formats are generated and managed in computational applications. Texts, movies, audio, maps, among others, are in the agenda of database administrators. How should such varied content be stored, processed and analyzed?

In order to extend the persistence options, new paradigms have been adopted in new databases called NoSQL (Not Only SQL). Tools such as Cassandra and Mongo-DB have been presented as solutions that address huge volumes of data with speed. distributed and low latency manner were used in social networking and other applications that require high levels of scale and availability. These tools generally apply to platforms such as Hadoop, which orchestrates the distributed file system and parallel processing to guarantee the performance of the implemented applications. Combining hardware and software specific appliances are processing systems in memory to provide performance in analytic queries. Vendors such as IBM, Oracle and SAP, are already available, allowing accelerating these transactions processing.

The discovery of unsuspected associations and summarization of data and visualization dashboards are on the agenda of organizations that aspire to become data driven. For this reason, software companies and the academic community are undertaking ongoing efforts in the search for models and more efficient and effective tools.

Organizations keep their data in silos or in isolation. Just as the systems of a city, the data produced in one application must be integrated with other, providing new perspectives in the context in which they were generated. However, new efforts are developed to add value from the data-driven, providing interaction between systems, to reduce the characterization of data silos in a Big Data Platform.

The integration of systems has been a trend, considering that the data can be enriched with new data to provide extra capacity for analysis and a better understanding of the processes of a company. However, to ensure systems integration, it is necessary to further studies on standards and reference models to ensure interoperability in distributed and open environments.

Another point is interoperability. This is the ability of a system to communicate or share data with other systems via interfaces or brokers without expending additional efforts for this activity. Contemporary organizations need interconnected information systems to encompass broader organizational contexts, and to reduce the technological limits. These demands are moving to more open, integrated, flexible, modular, federated, secure and transparent systems specifications. Models such as the RM-ODP (Reference Model - Open Distributed Processing) have been instantiated in projects to reach the needed solutions.

The opportunities to apply these concepts are numerous in finance, health, safety, manufacturing etc. McAfee and Brynjolfsson (2012) conducted studies and concluded that companies where BDA is effectively used are 5% more productive and 6% more profitable than their competitors - those numbers are a powerful argument for using this approach. In the current scenario, there are several examples of BDA available. Applications for Marketing, e-Commerce, e-Health, among others. In addition, there are barriers to the deployment of BDA, such as the costs involved, lack of C-level sponsorship, issues about hardware scalability and availability of software inside the organization. However, the greatest barrier is the lack of qualified human resources.

Nevertheless, to create value from the data requires a number of steps. Understanding each step is important to establish the organizational data architecture able to address problems in an agile and scalable way. These issues are embedded in the digital transformation. Cloud Computing, Mobile Systems, Analytics and Social Media are the objects of study on the business agenda.

These issues are interconnected and produce unlimited possibilities for innovative products with a direct impact on the quality of life. In an interdisciplinary way, we are creating products to monitor aspects of our health, what we do, what we eat etc. Data are collected on smart phones to be processed in the cloud environment.

The 4.0 Industry has attracted attention from researchers and generated advances with the Internet of Things and Analytics areas; systems called MES (Manufacture Execution System) and RPA (Robotic Processing Automation) are in the current agenda of CIOs and CEOs looking to position themselves as managers of Data-Driven organizations.

In conclusion, there is an emerging demand for professionals with a holistic vision, capable of using an interdisciplinary approach, to be able to apply analytical tools to the various contexts and scenarios of human activity. We believe that these skills are inherent to information systems professionals. Developing these constructs through academic research and by offering more effective curricula represent an important challenge for the information systems area.

14.3. The Challenge

Considering higher education in Brazil, professionals looking for skills and competencies to work in analytical systems can study Computer Science, Computer Engineering or Information Systems. Students should be able to act on issues such as requirements engineering, database modeling, management and monitoring of DBMS, following a closer approach to the analytical needs of organizations.

We understand that updating the academic curriculum is not trivial; the contents must be aligned with different stakeholders (the market, governments and schools) to build robust pedagogical projects that enable the proper training of undergraduate students.

Among the instruments that guide a Higher Education Institution in defining the contents for a computing curriculum is the National Curricular Guidelines (DCN), approved by the Ministry of Education in 2016 [MEC, 2016]. This defined as part of information systems skills: responsibility for collecting, storing and managing data to use in different areas, the creation of applications, systems and interfaces with programming and software engineering.

What has been observed over the years in this sub-area of computing is the movement of researchers and database professionals towards generating excellence in the use of database management systems and this has contributed to promoting essential technology in sensitive areas of Economy, Education, Health and Safety. However, as computer prices and sizes have fallen, new usage demands have been created, besides sophisticated collection needs, storage and processing in terms of speed, volume and high variety usage in the framework of Big Data.

Activities such as genetic analysis, sensor measurements and social networks rely on massive data storage to Exabytes with volumes of collection in Terabytes daily. This has led to a significant change in the database area. Traditional Relational DBMSs have difficulty archiving new levels of availability and scalability, maintaining the ACID properties. In response to this, industries and open user communities are creating new persistence mechanisms to support the new demands. Consequently, new paradigms, architectures, models and tools have been developed and now share space in the current market database, often coexisting in polyglot persistence applications. Moreover, IoT projects, another increasingly relevant topic, transversely involves hardware issues, network, database and computer architecture (IOTA, 2012).

These issues have amplified the difficulty in updating pedagogical projects regarding persistence issues. As previously mentioned, the DCNs for computing were proposed in 2012, and approved in 2016. This time lag impacts changes in the educational structure. Within the area of computing, a number of applications have since appeared, for example Waze which was unknown in Brazil before this time.

Therefore, defining key concepts to be taught on computing courses is a complex task. A change in menu goes through various approval bodies at the university. It must also be aligned with DCN, ENADE (National Examination of Students' Performance), SBC (Brazilian Computer Society) etc. besides meeting market needs and this all requires time.

Regarding Information Systems courses, the basic skills in database courses currently are: DBMS architecture, database design and Structure Query Language for manipulating datasets. This is not enough to meet current demands. Some IESs also include courses for programming database and/or trying to optimize the workload of courses, offering Business Intelligence disciplines and Multidimensional Modeling, Online Analytical Processing (OLAP) and Data Warehouse (DW).

A suggestion could be the treatment of these constructs as the basis for information system courses, building a deeper formation in order to generate skills, knowledge and professional attitudes. In the context of Bloom's taxonomy, this would include the adoption of verbs such as implement, develop, analyze, among others, to the contents of related disciplines.

Another challenge proposed in this text is the role of researchers to further the discussion about vocational training of Data Scientists, with regard to the basic concepts necessary to support projects in the Big Data Analytics platform [Breternitz et al. 2015]. Updates would be offered to Information Systems Pedagogic Projects or others, without extrapolating the course hours or affecting other contents.

This challenge can be addressed by the following:

- Discussion of initiatives to be proposed by researchers and that transcend the issue of including new issues into the pedagogical projects, such as the need to prepare lecturers to teach topics related to Big Data and Data Science, production of specialized papers and books and setting up Forums for constant discussion of the subject. This should generate inputs that can be quantified and thus used to assess the evolution of the challenge presented here.
- Definition of systematic research on the subject, to be held periodically, as a strategy to follow what has been researched in the area and relating this research focused on training people with skills in teaching subjects regarding Big Data and Data Science.
- Collective elaboration of a reference-syllabus necessary for working with data science, involving course coordinators, allowing, among other things, the mobilization of managers and the engagement of stakeholders.

14.4. Progress Evaluation

We believe that these topics can generate key performance indicators to follow the progress. Employment indicators and participation in forums can be measured, as well as the number of related publications.

14.5. Final remarks

Data science provides many opportunities as well as having positive and negative impacts on society. In a disruptive way, this wave has grown to allow us facilities to improve our quality of life. We have been innovating and reinventing how we live, at high speed. Hence, the field of Information Systems has a great role to play, as a part of an interdisciplinary world, with innumerable challenges, many of them still unknown.

References

Breternitz, V. J., Lopes, F. S. and Silva, L. A. (2015). Big Data Analytics: Education and Management of Data Scientists. CONTECSI - International Conference on Information Systems and Technology Management

Chen, P. P. S. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS)*, *1*(1), p. 9-36.

Clifford, S. (2012). Retail Frenzy: Prices on the Web Change Hourly. The New York Times, ed. 30.11.2012.

Davenport, T. H. (2014). Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Boston: Harvard Business School Publishing.

IoTA, Internet of Things Architecture (2012). Deliverable D1.3 - Updated reference model for IoT v1.5. European Commission within the Seventh Framework Programme.

Lohr, S. (2012). The Age of Big Data. The New York Times, ed. 11.02.2012. Available at: www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=1&scp=1&sq=Big%20Data&st=cse. Access in 09.05.2014.

McAfee, A; Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review. Available at: https://hbr.org/2012/10/big-data-the-management-revolution

MEC. Ministério da Educação (2016) Available at: http://portal.mec.gov.br/conselho-nacional-de-educacao/atos-normativos--sumulas-pareceres-e-resolucoes?id=12991 (In Portuguese)

Silberschatz, A., Korth, H. F., and Sudarshan, S. (2006). *Sistema de banco de dados*. Elsevier. (In Portuguese)

Simon, H. A. (1996). The Sciences of the Artificial. 3rd. ed. Massachusetts Institute of Technology.

Turban, E., Leidner, D., Mclean, E., and Wetherbe, J. (2010). *Tecnologia da Informação para Gestão-: Transformando os Negócios na Economia Digital*. Bookman. (In Portuguese)

Vaisman, A.; Zimányi, E. (2014). Data Warehouse Systems. Design and Implementation. Springer-Verlag Berlin Heidelberg.

Zikopoulos, P; et al. (2012). Harness the power of Big Data- The IBM Big Data Platform. Emeryville: McGraw-Hill Osborne Media.

Fábio Silva Lopes

CV: http://lattes.cnpq.br/2302666201616083

With a PhD degree in Environment Health, Fabio is a lecturer at the Faculdade de Computação e Informática of Universidade Presbiteriana Mackenzie (UPM). He is the coordinator of the Information Systems course. His research interests include: Information Systems, Big Data Analytics, Geographic Information Systems, and Software Engineering. He is a collaborating lecturer in The Professional Master's Program in Computer Engineering at IPT – Instituto de Pesquisas Tecnológicas de São Paulo. e-mail: flopes@mackenzie.br

Leandro Augusto da Silva

CV: http://lattes.cnpq.br/1396385111251741

Leandro Augusto da Silva is a Computer Engineer with a PhD in Systems Engineering from the School of Engineering of the University of São Paulo. He works on the following research topics: artificial neural networks, pattern recognition, data mining, machine learning and big data analytics. Leandro is currently working as an Adjunct Professor at the School of Computing and Informatics at Mackenzie University. e-mail: prof.leandro.augusto@mackenzie.br

Vivaldo José Breternitz

CV: http://lattes.cnpq.br/2865722030688852

PhD in Management, Vivaldo is a lecturer at the Faculdade de Computação e Informática (FCI) at the Universidade Presbiteriana Mackenzie (UPM). He is the coordinator of Internships. His research interests include Information Systems, Big Data Analytics, Enterprise Resource Planning and Software Engineering. He is a collaborator lecturer in the Professional Master's Program in Entrepreneurship at the São Paulo University (USP). e-mail: vjbreternitz@mackenzie.br