Business Intelligence & Analytics applied to Public Housing

Etienne Scholly

Business Intelligence & Analytics applied to Public Housing

Étienne Scholly[1,2]

[1] University of Lyon, Lyon 2, ERIC EA 3083
https://eric.ish-lyon.cnrs.fr
[2] BIAL-X
https://www.bial-x.com
etienne.scholly@bial-x.com

Abstract. Business Intelligence, with data warehouses, reporting and OnLine Analytical Processing (OLAP) are about twenty years old technologies, they are mastered and widely used in companies. Their goal is to collect, organize, store and analyse data to support decision-making. In parallel, there are many algorithms from Data Science for conducting advanced data analyses, including the ability to conduct predictive analyses. However, the reflection on the integration of Data Science methods into reporting or OLAP analysis is relatively incomplete, although there is a real demand from companies to integrate prediction into decision-making processes. In the meantime, with the rise of the Internet, the proliferation of multimedia data (sound, image, video, etc.), and the fast development of social networks, data has become massive, heterogeneous, of diverse and rapid varieties. The Big Data phenomenon challenges the process of data storage and analysis and creates new research problems. The PhD thesis is at the junction of these three main topics : Business Intelligence, Data Science and Big Data. The objective is to propose an approach, a framework and finally an architecture allowing prediction to be made in a decision-making process, but with a Big Data perspective.

Keywords: Business Intelligence · Data Science · Big Data

1 Introduction

Business intelligence (BI) refers to all the methods and tools used to collect, store, organize and analyze data. The objective is to provide decision-makers, such as business leaders, with an overview of the data as well as decision support, with the ultimate aim of improving the activity managed by these decision-makers [16]. BI is mainly based on the feeding and interrogation of a data warehouse [11]. The first phase is usually managed through ETLs (Extract - Transform - Load), which extract data from operational sources, transform them to meet upstream expectations, and store them in the data warehouse, which is very often a relational database. The second part consists in querying the data warehouse : it is mainly done through OnLine Analytical Processing (OLAP) and dashboards (reporting) [16].

Data Science allows to go further than the analyses proposed by BI. Data Science refers, in its broadest definition, to the extraction of knowledge from data. In this whole process, it is therefore necessary to clean, prepare, and analyze the data [15]. Data Science is at the intersection of several other disciplines, such as statistics, Artificial Intelligence (AI) and data visualization. One of the fields of study of AI is Machine Learning. It uses essentially statistical algorithms to give a computer the ability to "learn". Among these algorithms, we can distinguish two categories : unsupervised learning gathers observations into homogeneous groups ; and supervised learning, where one of the objectives is to create a model from a sample of data with a known membership class, and then test it with new data by trying to predict their class [14].

For a few years now, we hear about the "Big Data" phenomenon. Initially, this term referred to the amount of data available that is growing exponentially, but reducing this phenomenon to volume is too narrow. The first consensual definition of Big data is the "3 V's" : Volume, Variety and Velocity [7]. More recently, two V terms have been added to the definition of Big Data, thus creating the "5 V's" : Value and Veracity [4]. Traditional database storage and management systems have reached their limits and do not allow Big Data to be processed, forcing researchers and industrials to rethink all data processing methods, whether in BI or Data Science.

The goal of the PhD project is to unify BI with Data Science, in the general context of Big Data. It means that we want to be able to handle Big Data and enable users to run any kind of analyses, whether BI or Data Science ones. Indeed, we think that both types of analyses are needed and will co-exist. To this end, we will need a central storage area that can handle Big Data, and that can be queried by multiple applications downstream. BI analyses will mostly be conducted on structured data, while Data Science methods can process any kind of data. Data Science analyses will either be used as ad hoc analyses, or to feed the data warehouse with advanced indicators calculated through predictive methods (for instance).

The rest of the document is organized as follows. Section 2 presents the state of the art on Business Intelligence and Analytics. Section 3 explicits the two main problems of the PhD project. Section 4 announces results obtained so far and the work that is still to be done.

2 State of the art

The use of Data Science methods on a company's data is associated with the term *Business Analytics* (BA) [3]. Thus, by relying on the strengths of Data Science, BA makes it possible to carry out more advanced analyses than BI-specific analyses, and especially to "look forward" with supervised learning [12]. The ability to conduct predictive and prescriptive analyses opens up a wide range of new decision support applications [14].

Although the Data Science methods used by BA have been around for a long time, BA is still relatively new, particularly compared to BI. As a result, BA's definition is still rather vague, especially since the range of possibilities is very wide thanks to the diversity of existing Data Science methods. Some authors use the term *Business Intelligence & Analytics* (BI&A) to signify that the two domains, BI and BA, complement each other and form a whole [3]. Others suggest that BA is an extension of BI and tends to replace it by proposing new and more advanced analyses [14]. We believe the term is not essential because it is mainly marketing.

Authors have proposed BI&A architectures to meet the challenges of Big Data. Baars and Ereth [1] propose a BI&A platform that can manage Big Data. It relies on what they call *analytical atoms*, which can be seen as small, autonomous data warehouses ready for analysis. The atom's data has been historized, versioned, enriched, cleaned, etc. The idea is that there is an analytical atom for each incoming data source, and these atoms can be combined to create what the authors call virtual data warehouses. Gröger [8] proposes a very general data analysis platform, which is based on the *lambda architecture*. It is a standard for developing high-performance systems that can manage both batch and stream processing. Data is stored in a data lake, and numerous applications retrieve the data stored in the lake.

In addition to these BI&A architectures, a new approach to organizing and storing data is emerging : Data Lakes. This term was first introduced by Dixon [6], offering this alternative to datamarts, which are subsets of data warehouses. A data lake is a vast repository of raw data, of heterogeneous structures, and fed by several external sources. Data lakes rely on the *schema-on-read* property : data is stored in its raw format, and the schema is only specified when queried [13]. It is therefore mandatory to have an efficient metadata system in the data lake, which makes it possible to query data stored in the lake. A data lake without an efficient metadata system is called a data swamp, and is completely unusable [13]. From a technological point of view, most data lake implementations are based on Apache Hadoop, although new solutions are gradually emerging [13].

3 The PhD project

Our objective is to combine BI and BA in the general context of Big Data, by enabling users to conduct both types of analyses, separately or together, on Big Data. We consider that BA analyses can be performed as ad hoc studies (through clustering for example), and can also be used to consolidate BI analyses, particularly through advanced indicators, for example generated by predictive algorithms. Since most companies will always have "classic" BI analysis needs, we choose to keep a BI framework in our thinking, based on OLAP - if necessary - and dashboards.

Given the current state of the art in BI, BA and Big Data, we decide to draw inspiration from the work of Baars and Ereth [1], in particular their notion of an analytical atom, which we wish to expand. We are also inspired by the idea of a general data analysis platform such as the one introduced by Gröger [8]. Finally, we use the concept of data lakes, which will be at the heart of our approach.

More concretely, two types of data will feed the data lake : internal data, which is the company's data, and external data, collected from the Internet. Internal data will, in most cases, be structured data that does not present Big Data type problems. This structured internal data represents the core of the company's BI activity and is therefore traditionally managed by an ETL, which feeds a data warehouse, and reports are created from the data warehouse. External data, coming from very heterogeneous sources, can present Big Data issues. External data is the main source of BA analyses, although internal data can also be analyzed by these techniques. We believe that the main strength of external data lies in their capacity to be associated with the company's internal data, either by crossing them to have more variables or observations available, or by generating "advanced" indicators that can then be used in BI analyses.

We believe that the analytic atom concept, combined with the object notion [5], will help us propose a metadata system for a data lake efficient enough to manage any type of data, Big Data included. Although several studies have already been carried out on metadata systems for data lakes, and some of them have proven their efficiency [9, 10, 2], we believe we can offer a more complete metadata system that offers all the features we consider essential, and thus completely meets our expectations.

This constitutes our first problem. It is vital for our work to have a storage area in which to store all data, both internal and external. Various applications are powered by this storage area (in this case, the data lake) : data warehouse, predictive models, visualization, statistics, etc. The use of the lake is made possible thanks to a reliable metadata system, and the latter also helps users to better understand data and track its usage. In addition to the metadata system, the development of the entire data lake is necessary.

Once the development of the data lake and its metadata system is complete, we will have to feed it with data. This is the basis of our second problem. The PhD is conducted in partnership between the ERIC laboratory and BIAL-X through a CIFRE convention. Our work is anchored in the business issue of public housing, and more particularly in the study of the attractiveness of a social dwelling. Indeed, this is a key issue for BIAL-X because a significant part of the company's activity is dedicated to social landlords. The latter have data about their dwellings (number of rooms, surface area, energy category, construction date, etc.). However, finding information on the environment in which the dwelling is located is a more complex problem : this can be found in external data, collected from the Internet.

Combining external data with internal data can lead to the discovery of new and valuable information. For example, users can create "advanced indicators" with the help of analyses ran on external data to fine-tune indicators already existing in BI analyses. Even unstructured data can be processed : tweets to find the general opinion on a district, or pictures to determine if dwellings are similar, for instance. Thus, we would like to provide social landlords with the opportunity to simultaneously manage their activity through "classic" BI analyses, discover hidden insights in their data with BA analyses, and finally enrich BI analyses with advanced indicators such as the attractiveness of their dwellings. All these use cases would be derived from the data lake.

4 First results and future outcomes

We have started to work on the first issue presented in Section 3. This work was carried out with another doctoral student from the ERIC laboratory who works on textual data lakes. After an overview of the definitions of a data lake from the literature, we proposed our own definition of this concept. We then identified six key features that the metadata system of a data lake must provide, in our opinion, to be as robust as possible in addressing the Big Data issues and the schema-on-read approach. Comparing existing metadata systems, we showed that some works offer five out of six features, but none offers all six.

We proposed a metadata typology for data lakes. It is based on the object notion, which represents any set of homogeneous data [5], and the typology declines metadata into three categories. Intra-object metadata describe objects through versions, representations and various properties to name a few ; inter-object metadata explain how objects are linked together ; and global metadata facilitate and improve data analyses and the use of the data lake in general.

We also introduced a graph-based model of our metadata typology named MEDAL. An object is represented by a hypernode, and it contains nodes (representations and versions) connected via edges (updates and transformations). Hypernodes are linked through edges or hyperedges. Theoretically, MEDAL proposes all six key features identified.

However, we have not yet implemented MEDAL. This constitutes a technological challenge : we will have to identify which technologies we think are relevant to implement our data lake. There are many existing tools to implement a data lake, but after a quick review, we think that we might have to develop our own graph-based metadata management interface. At first sight, Apache Kafka seems to us to be a suitable way to implement our data lake, although other ways are still possible. Among our upcoming studies, a complete survey on existing tools will be carried out.

Once the data lake is implemented and functional, it will then have to be supplied with data. This constitutes the answer to the second issue presented in Section 3. We will insert into the data lake the social landlord's data, which mainly contains the properties of social dwellings, as well as external data retrieved from the Internet, mainly from open data. The latter are the ones that will have the most Big Data problems, especially in terms of data variety. Our ultimate objective will be to exploit the capabilities of our metadata system to have both raw and reworked data in the lake, to feed various applications (including a data warehouse and predictive models), to try to fine-tune the attractiveness of a home, and to enable social landlords to better manage their activity by comparing their data to external information.

References

1. Baars, H., Ereth, J.: From data warehouses to analytical atoms-the internet of things as a centrifugal force in business intelligence and analytics. In: 24th European Conference on Information Systems (ECIS), Istanbul, Turkey. p. Research-Paper3 (2016)
2. Beheshti, A., Benatallah, B., Nouri, R., Chhieng, V.M., Xiong, H., Zhao, X.: CoreDB: a Data Lake Service. In: 2017 ACM on Conference on Information and Knowledge Management (CIKM 2017), Singapore, Singapore. pp. 2451–2454. ACM (November 2017). https://doi.org/10.1145/3132847.3133171
3. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS quarterly pp. 1165–1188 (2012)
4. Chen, M., Mao, S., Liu, Y.: Big data: A survey. Mobile networks and applications **19**(2), 171–209 (2014)
5. Diamantini, C., Giudice, P.L., Musarella, L., Potena, D., Storti, E., Ursino, D.: A New Metadata Model to Uniformly Handle Heterogeneous Data Lake Sources. In: New Trends in Databases and Information Systems - ADBIS 2018 Short Papers and Workshop, Budapest, Hungary. pp. 165–177 (September 2018). https://doi.org/10.1007/978-3-030-00063-9_17
6. Dixon, J.: Pentaho, Hadoop, and Data Lakes. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (October 2010)
7. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management **35**(2), 137–144 (2015)
8. Gröger, C.: Building an industry 4.0 analytics platform. Datenbank-Spektrum **18**(1), 5–14 (2018)
9. Halevy, A.Y., Korn, F., Noy, N.F., Olston, C., Polyzotis, N., Roy, S., Whang, S.E.: Goods: Organizing Google's Datasets. In: Proceedings of the 2016 International Conference on Management of Data (SIGMOD 2016), San Francisco, CA, USA. pp. 795–806 (june 2016). https://doi.org/10.1145/2882903.2903730
10. Hellerstein, J.M., Sreekanti, V., Gonzalez, J.E., Dalton, J., Dey, A., Nag, S., Ramachandran, K., Arora, S., Bhattacharyya, A., Das, S., Donsky, M., Fierro, G., She, C., Steinbach, C., Subramanian, V., Sun, E.: Ground: A Data Context Service. In: 8th Biennial Conference on Innovative Data Systems Research (CIDR 2017), Chaminade, CA, USA (January 2017), http://cidrdb.org/cidr2017/papers/p111-hellerstein-cidr17.pdf
11. Inmon, W.H.: Building the Data Warehouse. John Wiley & Sons (1996)
12. Larson, D., Chang, V.: A review and future direction of agile, business intelligence, analytics and data science. International Journal of Information Management **36**(5), 700–710 (2016)
13. Miloslavskaya, N., Tolstoy, A.: Big Data, Fast Data and Data Lake Concepts. In: 7th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2016), NY, USA. Procedia Computer Science, vol. 88, pp. 1–6 (December 2016). https://doi.org/10.1016/j.procs.2016.07.439
14. Mortenson, M.J., Doherty, N.F., Robinson, S.: Operational research from taylorism to terabytes: A research agenda for the analytics age. European Journal of Operational Research **241**(3), 583–595 (2015)
15. Shmueli, G., Koppius, O.R.: Predictive analytics in information systems research. MIS quarterly pp. 553–572 (2011)
16. Watson, H.J., Wixom, B.H.: The current state of business intelligence. Computer **40**(9), 96–99 (2007)