Taming Corporate Data

Iacopo Ennio Inghirami

University of Milano-Bicocca, Italy
iacopo.inghirami@unimib.it

Abstract. Today, every company is faced with a large amount of data that reaches it through multiple communication channels. To "tame" these data, the company (1) needs to carefully consider the data it receives to decide what to do with it; (2) must evaluate which procedures and tools to use; (3) must invest in Human Resources to manage data and information. We need new facilities to implement, we need new professionals, the Chief Data Officer and the Data Scientist, we need to increase final user knowledge by means of Data Literacy.

Keywords: Accounting Information Systems, Marketing Information System, Data Explosion, Data Visualization, Chief Data Officer, Data Scientist, Data Literacy

1 Introduction

Modern management is based on many concepts, including the speed of decisions. In a rapidly changing environment characterized by a high level of competitiveness, it is of crucial importance to be able to make decisions that are quick and yet as documented as possible. While in the past data and information were scarce, today the amount of data that oppresses companies is relevant.

Company Information Systems are facing a great challenge: the explosion in the amount of data and information. Arne Holst, a researcher at Statista, a company that deals with statistics relating to the world of communication, recently stated that:

> "The total amount of data created, captured, copied, and consumed globally is forecast to increase rapidly, reaching 64.2 zettabytes in 2020. Over the next five years up to 2025, global data creation is projected to grow to more than 180 zettabytes. In 2020, the amount of data created and replicated reached a new high. The growth was higher than previously expected caused by the increased demand due to the COVID-19 pandemic, as more people worked and learned from home and used home entertainment options more often." [9]

The main response to this phenomenon has been to offer companies greater data storage capacities and tools capable of managing this enormous amount of data, i.e., those tools that are used for "Big Data Analysis". In essence, Big Data Analysis tools have been proposed as a panacea to manage and understand all data. In many

environments, Big Data Analysis can provide interesting perspectives that can really guide business strategies. However, in our opinion, it makes no sense to deal with data relating to Accounting and Sales with Big Data tools.

> "Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value." [3]

Accounting and Sales managers are perfectly familiar with the data that feeds their respective systems. We are indeed faced with data that is neither High Volume nor Variety. A Chief Financial Officer (CFO) will be able to describe in detail the structure and contents of the data flows that feed the Accounting Information Systems.

In the same way, a Chief Sales Officer (CSO) can describe the structure, contents and correlations that link the data of the flows that feed the Marketing Information System (MKIS). Just as it will be able to predict certain customer behaviours without resorting to sophisticated Data Mining techniques or heavy Big Data Analytics procedures.

The point is that all the data relating to accounting comes from well-structured flows and from known sources. The case of Marketing data is different: in addition to structured data flows from known sources, there may be unstructured data sources from different sources. Furthermore, we may notice that in Marketing, unstructured data prevails over structured data.

It is necessary to rethink the Information Systems and redesign them so that they are ready to face the new challenges, that is to be able to manage structured and unstructured data. In the present work we will illustrate the three trends that will allow Information Systems to be revolutionized:

1. The Data Lakes, which will replace the current Data Warehouses.
2. The inclusion in the company of new human resources: the Chief Data Officer and the Data Scientists
3. The need to develop adequate Data Literacy in the company.

How can companies cope with it and "tame" the staggering quantities they receive daily? With this paper we will try to answer these three questions:

RQ1: How should the tools used in the company evolve?
RQ2: How should data processing procedures change?
RQ3: How should organizational structures evolve to face the challenges related to data and information management?

2 Rethinking Technological Structures.

2.1 The Classic Data Warehouse Model.

We can identify at least three main categories of data and information: Raw Data, Processed Data and Information.

- Raw Data is the data that arrives through the company's communication channels.
- Processed Data are structured, organized, cleaned up data which is stored in special Data Warehouse Servers and obtained from Raw Data through the Extraction, Transformation and Loading (ETL) procedures.
- The BI stage data and information are the data and information obtained from end users using Business Intelligence tools.

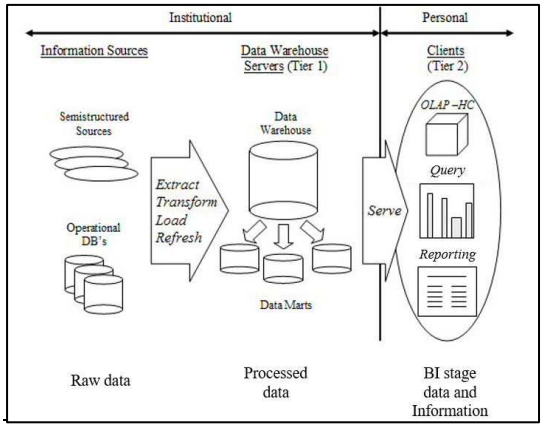

Fig. 1. – From raw data to information (classic Data Warehouse Model)

The management of Raw Data and the procedures that transform such data into Processed Data are managed by the company's IT structures [10]. The management of data and information in Business Intelligence systems is instead delegated to the end users (see fig. 1).

2.2 The Data Lake Model.

Data Lake (DL), as a relatively new concept, is defined in both academic community and industrial world [16]. All the existing definitions respect the idea that a DL is a repository storing raw data in their native format.
We will have data from Web Logs, social media, Corporate Systems (OLTP, ERP, CRM, etc.), various documents, emails, data from IoT devices and other Cloud devices. Yet, different definitions have different emphases [16]. Regarding input, [5] introduces that the input of a DL is the data within an enterprise. Regarding process, [13] emphasizes that there is no process during the ingestion phase and [2, 6, 13, 15] introduce that

data will be processed upon usage. Regarding architecture, [5] presents that DLs are based on an architecture with low-cost technologies. Regarding governance, metadata management is emphasized in [6]. And regarding users, [12] presents that data scientists and statisticians are DL users.

Basically, Data Lakes are the evolution of Data Warehouses: while in the latter the data must be structured, the DLs are able to manage any form of data, structured or not. Another feature that almost all DLs have in common is that of being in the Cloud, to be scalable and inexpensive.

The main procedures concerning the DLs are:

1. DATA INGESTION: Extracts data from a variety of sources and loads it into the lake.
2. DATA PROCESSING: Runs transformation routines and algorithms on raw data.
3. STORAGE: Stores vast quantities of data in a range of formats.
4. ANALYTIC SERVICES: Allows users to analyse processed data for a variety of use cases.
5. SECURITY & GOVERNANCE: Ensures the availability, usability, and integrity of data.

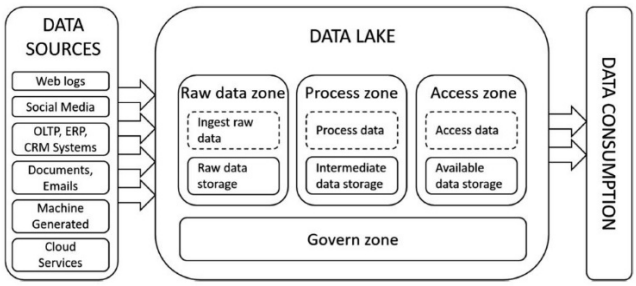

Fig. 2. – From raw data to information (Data Lake Model) [16]

The main idea of DLs is to ingest raw data without process and process data upon usage [16]. Therefore, Data Lakes keep all the information and have a good flexibility. Nevertheless, DLs, which contain a lot of datasets without explicit models or descriptions can easily become invisible, incomprehensible and inaccessible. So that it is mandatory to set up a Metadata Management System for DL [17]. In fact, the importance of metadata has been emphasized in many papers [1, 6]. The first problem that needs to be solved is the content of the metadata.

3 Rethinking Organizational Structures.

3.1 An emerging role: the Chief Data Officer.

The chief data officer has a significant measure of business responsibility for determining what kinds of information the enterprise will choose to capture, retain and exploit

and for what purposes. A chief data officer's purpose is to connect the technological results to the needed business results. Various other roles entail understanding the business value. It means using data to derive business outcome. Some responsibilities include the governance, advising & monitoring enterprise data. In terms of operations, it means enabling data usability along with efficiency and availability. They must innovate which means driving the business towards digital transformation innovation, cost reduction, and revenue generation. Their role is also to provide supporting analytics with reports on products, customers, operations, and markets. [20, 14, 11]
In many ways the job of the Chief Data Officer is like that of the Controller. The Controller is devoted to enable the four characteristics of management control [8]: repair, internal transparency, global transparency and flexibility related to the new budgeting practices in one global paper company. Findings The implementation of rolling forecasting was a major attempt at 'repair' to remedy the incompleteness of accounting information, which made controllers experts in producing and delivering more realistic forward-looking information in the organization. The increasing internal and global transparency of new budgetary practices enabled controllers at various levels of organization to develop new competences, which helped controller network to build a holistic view of the totality of control and supply more relevant information in organization.

3.2 A new profession: the Data Scientist

More than anything, what data scientists do is make discoveries while swimming in data. It's their preferred method of navigating the world around them. At ease in the digital realm, they can bring structure to large quantities of formless data and make analysis possible. They identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set. In a competitive landscape where challenges keep changing and data never stop flowing, data scientists help decision makers shift from ad hoc analysis to an ongoing conversation with data. [18, 3]

4 Widespread capabilities: Data Literacy

Whilst the fields of machine learning, data analysis and visualization are rich, there is surprisingly little research into the human component, particularly as it applies to more complex data. What are the competences that people must acquire to learn from and solve problems with data? What new skills must humans learn in order to both design, interpret and critique complex data analysis and visualisation?

The term "data literacy" is used to broadly describe the set of abilities around the use of data as part of everyday thinking and reasoning for solving real-world problems. Data literacy is increasingly considered to be a life skill, as daily interactions with data become ever more commonplace and individuals more frequently make judgments from data and make decisions regarding the use of their own personal data [19]

The problem of Data Literacy is particularly relevant for two reasons: on the one hand, companies require human resources capable of interacting profitably with

systems, managing to make the most of them. On the other hand, it is necessary for the academic world to adapt its courses and provide its students with all the Data Literacy that allows them to take advantage of new capabilities of information systems.

Furthermore, as we have found, the DL systems are dedicated to data scientists and statisticians and therefore a further step forward is necessary that also allows end users to take advantage of new technologies. This means improving current DL systems or adding interfaces to these systems that facilitate their use.

5 Conclusions

We dream of an agile but robust system, able to support the multiple needs of end users. Unfortunately, the increasingly complex environment pushes us to find even more complex solutions. However, it is necessary to continue to seek a certain ease of use, so that information systems do not remain the prerogative of specialists.

With this paper we had try to answer these three questions:

RQ1: How should the tools used in the company evolve?

RA1: It is necessary to replace the old systems, the Data Warehouses, with the more modern Data Lakes, which can manage any type of corporate information.

RQ2: How should data processing procedures change?

RA2: The systems incorporate data and information and process them on demand using metadata. These processed data will be used to create tables, graphs and anything else that can be used to manage the company.

RQ3: How should organizational structures evolve to face the challenges related to data and information management?

RA3: From the point of view of Human Resources, we must bring two new figures into the company: the Chief Data Officer (CFO) and the Data Scientist. It is also necessary to introduce the concept of Data Literacy in the company, which allows all users to create tables and graphs to illustrate their information.

References

1. Alserafi, A., Abell´o, A., Romero, O., Calders, T.: Towards information profiling: data lake content metadata management. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 178–185. IEEE (2016)
2. Campbell, C.: Top five differences between data lakes and data warehouse, January 2015. https://www.blue-granite.com/blog/bid/402596/top-five-differencesbetween-data-lakes-and-data-warehouses (2015)
3. Davenport TH, Patil DJ. Data scientist. Harvard business review. 2012 Oct;90(5):70-6. (2012)

4. De Mauro A, Greco M, Grimaldi M.: A formal definition of Big Data based on its essential features, Library Review, (2016)
5. Fang, H.: Managing data lakes in big data era: what's a data lake and why has it became popular in data management ecosystem. In: 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 820–824. IEEE (2015)
6. Hai, R., Geisler, S., Quix, C.: Constance: an intelligent data lake system. In: Proceedings of the 2016 International Conference on Management of Data, pp. 2097–2100. ACM (2016)
7. hang H, Lee Y, Wang R, Huang W. Chief data officer appointment and origin: A theoretical perspective. (2017)
8. Henttu-Aho T. Enabling characteristics of new budgeting practice and the role of controller. Qualitative Research in Accounting & Management. 2016 Apr 18. (2016)
9. Holst, A.: Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025, https://www.statista.com/statistics/871513/worldwide-data-created/, (2021)
10. Inghirami IE. Reshaping strategic management accounting systems. In DSS 2.0–Supporting Decision Making with New Technologies 2014 (pp. 495-506). IOS Press. (2014)
11. Lee Y, Madnick SE, Wang RY, Wang F, Zhang H. A cubic framework for the chief data officer: Succeeding in a world of big data. (2014)
12. Madera, C., Laurent, A.: The next information architecture evolution: the data lake wave. In: Proceedings of the 8th International Conference on Management of Digital EcoSystems, pp. 174–180. ACM (2016)
13. Miloslavskaya, N., Tolstoy, A.: Big data, fast data and data lake concepts. Procedia Comput. Sci. 88, 300–305 (2016)
14. Nie Y, Talburt J, Dagtas S, Feng T. The influence of chief data officer presence on firm performance: does firm size matter?. Industrial Management & Data Systems. 2019 Apr 8. (2019)
15. Piatetsky-Shapiro, G.: Data lake vs data warehouse: key differences, September 2015. https://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-keydifferences.html (2015)
16. Ravat, F., Zhao, Y.: Data Lakes: Trends and Perspectives. International Conference on Database and Expert Systems Applications (DEXA 2019), Aug 2019, Linz, Austria. pp.304-313. (2019)
17. Ravat, F., Zhao, Y.: Metadata management for data lakes. In: East European Conference on Advances in Databases and Information Systems. Springer (2019)
18. Waller MA, Fawcett SE. Click here for a data scientist: Big data, predictive analytics, and theory development in the era of a maker movement supply chain. (2013)
19. Wolff A, Gooch D, Montaner JJ, Rashid U, Kortuem G. Creating an understanding of data literacy for a data-driven society. The Journal of Community Informatics. 2016 Aug 9;12(3). (2016)
20. Xu F, Zhan H, Huang W, Xin (Robert) Luo, Xu D. The Value of Chief Data Officer presence on Firm Performance. InPACIS 2016 Jun 27 (p. 213). (2016)