To Puzzle Out

Sónia Patrícia F. C. B. Quaresma Gonçalves

Information Systems and Computing Department
National Statistical Institute
Portugal

sonia.quaresma@ine.pt

Abstract: Since the beginning of times we've been trying to comprehend our reality, making sense of what happens and gathering information about it. Understanding our environment gives us the power to change it, adjust it or just to act in accord with its latest development.

In a complex and ever-changing world multiple aspects have to be accounted for, demographics and finance, environment and agriculture, economics and research… National Statistical Institutes pay attention to all these areas, gathering information about each and every one of them.

To apprehend the reality is our aspiration at National Statistical Institutes and to do it we use every possible tool at our disposal. The aim of this paper is to present one of the most innovative tools in NSIs the Data Warehouses. We'll talk about the evolution in data manipulation made possible by the technological advances in data collection which resulted in an increasing demand for more powerful ways to deal with the data, thus making DWs necessary.

Analysing DWs' strengths in the business world, we'll make clear their advantages for NSIs. Beyond that we shall see why they're so well adjusted for an NSIs' common needs of data dissemination and derived statistics production.

Even if DWs were not specially developed with statistical objectives, NSIs' purpose, when data is collected is not only to gather information but furthermore to build Knowledge about the world. In order to do that, we have to be able to relate information to one another. Linking information from different areas we'll give us different perspectives on the problems and an extraordinary new insight. And that's precisely what data warehouses do, so that we can slowly begin to puzzle out our world!

Keywords: Data Warehouses

1. Introduction

This paper is structured in four simple sections. Following this introduction second two makes a brief historical incursion in the databases development to clarify what are data warehouses, their purposes and major differences regarding relational databases. Section three focuses on the multidimensionality inherent to the data warehouse philosophy and centres its attention in the advantages it brings when working with statistical information. We also present

some examples to make clearer the concept of multidimensionality so that its benefits to data quality will emerge naturally. In section four, we present our conclusions.

2. What are Data Warehouses?

Humans have always tried to make sense of what surrounds them, gathering information in such a way that it could be passed to future generations. The accumulation of different data collected in similar circumstances spontaneously gave birth to a statistical approach to life.

By observing the seasons throughout decades and following the rain patterns our ancestors were able to establish the best harvesting cycles. Since then, our strategies to act upon our surroundings have evolved but always based on information collected from our environment, because that is what enables us to understand our world.

2.1 Historical overview

Since 442 B.C. a magistrate of high rank in the ancient roman republic occupied a position called *censura*, whose duties were to register the citizens and their property, to keep the public morals and to administer the finances of the state, namely superintending the public buildings and the erection of all new public works [1].

Each citizen had to give an account of himself, of his family, and of his property upon oath and was taxed one per thousand upon the property entered in the books of the censors. The lists of the persons and their ages were later used to call valid men to integrate the roman army. Being able to know how many *capita* (heads) could be called to serve as soldiers, at any given moment, was very important, and this counting of heads is still imperative today. In fact the census is in most countries one of their biggest statistical operations, even if it is derived from administrative information. Likewise the assessment of state finances is vital and, in both cases, information is disseminated in this day and age by the National Statistical Institutes, as it was in roman times published in the *Tabulae Censoriae*.

Historically this knowledge was preserved in the written form in books, from the ancient library of Alexandria in the 3rd century B.C. to contemporary public information repositories such as the Library of Congress, and until some 50 years ago that was the only way to access the information: read the book.

As human activities grew more complex bigger amounts of information were needed to take any decision. With the advent of computers books were compressed in electronic files but that was still not enough. In 1968 [2] to support the Apollo lunar missions, IBM [3] developed the System/360 Model 85 contributing to the successful landing of the Apollo 11. This information management system (IMS [4]) marked the birth of databases and over 300 patents directly or indirectly related to the System/360 were issued between 1964 and the end of the decade.

These databases were just tables with records. Anything that could be coded would, so that the entire record occupied even less.

Rules to codify the table fields were proposed by Codd [5], as early as 1970, it was the beginning of relational databases, and in 1976 Chen [6] presented his well known entity/relationship model with very specific techniques to relate tables to each other in such a way that more complex systems could be modeled. The resulting easiness of data capture led to the present situation in which every transaction we perform, from a simple acquisition to a flight is recorded in several databases throughout the world.

Possessing so much data made the companies aware of its value not only to better serve their clients but also to promote additional business with them if they could predict or discover their needs and wants. Excellency in organizations [7] is measured through their tangible achievements in what they do, how they do it and what they can accomplish. If the organizations were able to reorganize all their data in different perspectives, they would have an added value product without further work in information gathering.

The only problem was that the databases, into which this data was entered, were the so-called online transactional processing databases (OLTP [8]), which complied with all Codd rules about eliminating redundancy and comprising data. This means data insertion was optimized but not data analysis or extraction, to feed other analytical systems.

Getting information from these OLTP databases was hard and very slow, not only because of the way the data was physically arranged but also because the perspective in which it was collected was not the one best suited for analysis. For example, when we registered a sale with all its small details the record had all the information of that specific sale, but that is not the perspective important to the supervisors or managers. What they need are summaries of all the sales and not the specificities of each one of them.

Even worse, the problem of the analysis delay appeared suddenly: gathering data was faster everyday, but its analysis was growing slower with its increase.

More and more companies were trying to use their business data for analysis and becoming unsatisfied with the results achieved. As in any other fields, industry pressure soon produced results and just 20 years after the generalized use of relational databases did Inmon [9] propose a different kind of databases, a digital information repository that would be called a data warehouse. The concept of online analytical processing databases (OLAP [10] [11]) was born.

3. Multidimensionality

The purpose of OLAP databases was to keep the data in such a format that its extraction would be easier than from traditional OLTP systems [12], but also to rearrange it or, in other words, transform it from data into information [13].

The first problem, optimizing data extraction, consisted in inserting all the redundancy that Codd and the developers of OLTPs were so eager to get rid of.

As to the second problem, turning data into information, in most cases it amounted to the construction of several summary tables.

Once this was accomplished getting the added value product that information is to companies was now easier than ever thanks to the use of data warehouses and OLAP systems. With this kind of information, organizations were able to build knowledge, understanding their customer and building awareness to their crescent needs. The effort was then redirected to have not one added value product, resulting from the collected information, but several different products or perspectives upon the same information.

Soon was realized that contemplating distinct angles of a problem was just a matter of changing the dimensions of the analysis [14] [15]. In fact, the same sale could be studied in a temporal perspective, like sales in a quarter or geographical, like sales of a particular store in Lisbon, or even product type, as in sales of red bricks.

The possible points of view of a same problem were virtually endless. This led to the use of hyperspace or multidimensional as terms applicable to database technologies. Of course visualizing a space with so many dimensions or hyperspace is very difficult to a human, to say the least. The theory can nevertheless be easily understood by reducing those dimensions to, let's say, three in which case we'd have a cube to represent our problem.

If we picture Rubik's cube, with a different colour in each face, it contains in itself 27 other small cubes, each one of them at a unique crossing of the three dimensions at stake [16].

Figure 1 - 3D Cube manipulation

How does this multidimensionality apply to statistics? In every single thematic we approach at National Statistical Institutes, Geography and Time are always present [17]. All events happen somewhere in space and sometime. These characteristics are the starting point to the examination of any phenomenon.

Let's say we want to study Nuptiality. When two persons decide to share their lives and cohabit, they fill some forms declaring where they are going to live, the declaration date, and how did they join their lives: through a religious ceremony? In a civil marriage? Or they just moved-in?

More information is required by the administrative office, but let us focus on the three dimensions we already have here: Geography, Time and Type.

The Type of Nuptial bond has three possible values, as we have seen: Civil, Religious or Moved-In.

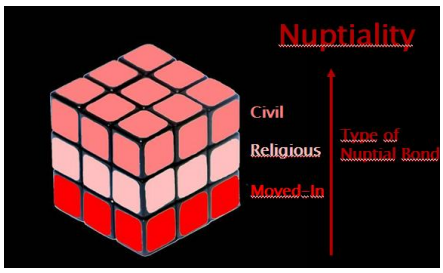

Figure 2 - Type of Nuptial Bond

When we speak about Time we can choose several different periods. Distinct aspects of a problem can be better understood by considering different time periods, like month, year, or season [18]. If we are interested in discovering if people prefer spring over winter to get married we should focus on seasons. On the other hand, if we are trying to discover if the number of marriages has been decreasing lately, we could choose the latest three years.

Finally when we think about Geography it is also important to suit the regions we choose to our particular study. For example, if we are studying the marriage behaviour in the biggest cities in central Portugal, we might choose Lisbon, Sintra and Cascais.

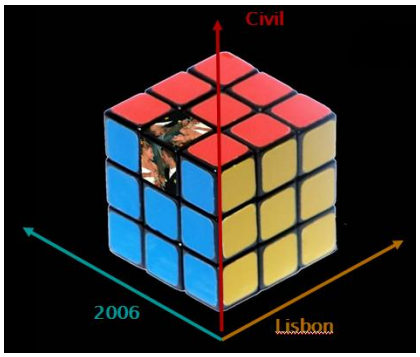

Figure 3 - Civil marriage celebrated in 2006 in Libon

As we've stated earlier every position in our Rubik cube will represent a combination of choices in the three dimensions described. This aggregation of data has not only the benefit of simplifying the study of a phenomenon at a macro level but also preserves the privacy at the individual level while allowing accurate models to be constructed [19].

Another benefit of the summary tables of data warehouses and its multidimensional treatment of data is the ability it gives us to build reports much quicker than before with biggest amounts of information [20].

Most data at National Statistical Institutes has a geographical and a time dimension. So if we were studying live births we could aggregate all the births by year, geographical region and type of birth.

Figure 4 - Live Births Cube

Using the shared dimensions of Geography and Time we can relate the information we receive from numerous sources or even from several surveys and combine it in data warehouses [21] building meta cubes.

In this case instead of three dimensions in the Live Births Cube we could have four. The first three would be Geography, Time and Type of Birth, as we've already suggested but the fourth could be the entire Nuptiality cube linked to the new cube by its geographical regions and by the periods chosen that would have to be the same.

In some cases the information collected is not treated at the same level in the several cubes we possess and we could have to aggregate it some more in one of the cubes. For example, in spite of having the Nuptiality cube organized by year the department responsible for the construction of the Live Births cube could have chosen monthly periods. In this situation we would have to summarise the Live Births to a yearly basis before we could combine the 2 cubes [22].

The effort of harmonization of the geographical dimension can be strenuous but causing quite a few cubes coalesce can help us spot hitches regarding the data quality and identify problematic areas or sources [23].

As important as the diagnostic of our data quality is that meta cubes enable us to extend our knowledge on the subject permitting us to answer more complex questions like: "Most children are born within civil or religious marriage?" or "How long do people usually wait between marriage and the first birth?". This can help us confirm previous ideas we already had on the field but also discover recent patterns that had not previously been spotted.

4. Conclusions

We briefly presented the evolution of the treatment of information, since the birth of OLTP to the emergence of OLAP databases, followed by the description of concepts such as multidimensional cubes or hypercubes (cubes in hyperspace). Examples were given to clarify those concepts.

The advantage of hypercubes for NSIs is that in most cases we can hold down two dimensions: time and geography while varying the others. In this way, we're able to relate subjects that otherwise would be disconnected. Furthermore, to link two hypercubes its necessary that they share at least one dimension and as we've seen in NSIs they usually have two in common turning multidimensional cubes into a technology particularly well adjusted to NSIs, even if DWs were not specially developed with statistical objectives.

Use the information collected in the best possible way and build Knowledge about the world with it is our aim and gets us closer to puzzling out our world.

5. References

[1] Cicero: de Leg bus II, 3 (On the Laws). Approximately 44 B.C.

[2] IBM http://www-1.ibm.com/ibm/history/exhibits/space/spaceskylab.html (1968). Acedido em 14/07/2004

[3] W. I. Stanley, H. F. Hertel: *Statistics gathering and simulation for the Apollo real-time operating system*. IBM Systems Journal: Volume 7 Number 2 (1968).

[4] IBM http://www-03.ibm.com/servers/eserver/zseries/timeline/1960s.html Acedido em 09/04/2007.

[5] Codd E. F.: *A Relational Model of Data for Large Shared Data Banks*, in Communications of the ACM,Vol. 13, No. 6, pp. 377-87. (1970).

[6] Chen P. P.: *The Entity-Relationship Model - Toward a Unified View of Data.* ACM Transactions on Database Systems (TODS), Vo. 1 No. 1, pp. 9-36. (1976).

[7] EFQM: *Introducing Excellence.* European Foundation for Quality Management. (2003).

[8] Tsou & Fischer: *Decomposition of a Relation Scheme into Boyce-Codd Normal Form.* SIGACTN. (1982).

[9] Inmon W. H.: *Building the Data Warehouse*. Weinheim: Wiley. (1993).

[10] Bulos D.: *OLAP Database Design: A New Dimension*. Database Programming and Design. (1996).

[11] OLAP Council. OLAP AND OLAP Server Definitions. 1997.

[12] *A Survey of Logical Models for OLAP Databases*. SIGMOD Record. (1999).

[13] Buzydlowski, J., Song, I., & Hassell, L.: *A Framewok for Object-Oriented On-line Analytical Processing.* Proceedings of the Annual DOLAP Conference, USA, ACM 1-58113-120-8/98/1. (1999).

[14] Pedersen T.B and Jensen C. S.*: Multidimensional Data Modeling for Complex Data.* ICDE, 336 – 345. (1999).

[15] Nguyen T. B., Tjoa A. M. & Wagner R.: *An Object Oriented Multidimensional Data Model for OLAP*. Web-Age Information Management, 69 - 82. (2000).

[16] Lechtenborger J. and Vossen G.: *Multidimensional Normal Forms for Data Warehouse Design*. Elsevier Science. (2002).

[17] Sindoni G. and Tininini L.: *A Statistical Web Warehouse System.* International Conference on Quality in Official Statistics. Mainz, Germany. (2004).

[18] Gonçalves S.: *A Brave New World*. International Conference on Quality in Official Statistics. Mainz, Germany. (2004).

[19] Agrawal R.: *Next Frontier.* Microsoft Search Labs. SIGKDD International Conference. Philadelphia. (2006).

[20] Chiepa A., Palma A. and Zindato D..: *DataWarehousing Population Census:Reporting on Data during the Processing and Macrodata Validation Phases*. International Conference on Quality in Official Statistics. Cardiff, Wales. (2006).

[21] Winkler W.: *Data Quality: Automated Edit/Imputation and Record Linkage*. International Conference on Quality in Official Statistics. Cardiff, Wales. (2006).

[22] Gonçalves S.: *Easy Does It!* International Conference on Quality in Official Statistics. Cardiff, Wales. (2006).

[23] Dehn K. and Nielsen T..: *Dnaish Statistics of Education – on the road to Nirvana*. International Conference on Quality in Official Statistics. Cardiff, Wales. (2006).