Integrating micro-databases for statistical purposes

Paula Menezes[1] and Luís D'Aguiar[2]

Introduction

Data are critically important in making well-informed decisions. Poor quality data or, *a fortiori*, lack of data can lead to inefficient allocation of resources and imposes high costs on Society. The strategy defined by the *Banco de Portugal* (hereinafter referred as "the Bank") to deal with the challenge of maintaining its statistics relevant to the users in a shifting and more demanding environment, while attending to the need to keep the respondents' reporting burden at an acceptable level, was to enhance the overall efficiency of the statistical framework by further exploring the largely unused statistical potential of already existing data sources (including the available administrative micro-databases). This approach has been facilitated by the advancements in information systems and technologies (IS/IT), network and communication protocols, database systems and multidimensional analytical systems, which have removed most of the potential shortcomings of having to deal with the huge amounts of data usually associated with the handling of micro-databases.

However, expanding the sheer amount of data available, *per se*, does not necessarily correlate with its value – *"not everything that can be counted counts"* (citation accredited to Albert Einstein). Indeed, there is a need for tools that enable rapid data exploration, permitting multidimensional analysis and cross-reference of multiple sources of information with different granularity. We this in mind, the Bank decided, back in 2008, to lay the foundations of a reference framework for the planning and implementation of IS/IT projects in the Statistics Department, so that the solutions developed within the scope of the different statistical domains could contribute to the incremental construction of a single, coherent and integrated repository of information.

Aiming at micro-data integration

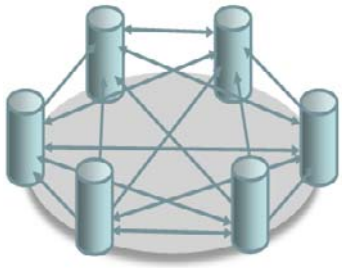
The situation prevailing at that juncture could be described by the existence of multiple isolated systems, in some cases legacy systems, built with the sole purpose of addressing specific needs identified within the restricted scope of each development project. The statistical information was distributed across multiple data "silos" of limited accessibility (see Figure 1. below), and the use of each information system was almost entirely vertical, that is, confined to a business area.

[1] Statistics Department, Banco de Portugal, pamenezes@bportugal.pt.

[2] Statistics Department, Banco de Portugal, laguiar@bportugal.pt.

From the standpoint of the access of the various applications to data from other systems, for the purpose of cross-checking or integrating information, the solutions that have been created consist in extracting and transmitting data from each system to the others, according to the different needs and in varied formats.

Direct integration of ISs — Figure 1



This type of integration model results in a plurality of interfaces, often with poor support; the processes of extraction, transformation and loading occur redundantly and with increased maintenance costs. Another problem associated with this form of integration lies in the difficulty in ensuring data coherence, and in guaranteeing that data updates are reflected equally in the various applications or that reference data are consistent in all databases.

Under this application architecture there were not sufficient resources to effectively respond to the growing need of integration and sharing of information. This problem manifested itself in different ways: from the point of view of those who use the information, by the difficulty of access and cross-check different databases, with diverse structures and formats, or simply by the lack of means to access data directly; from the statistical production side, by the effort involved in establishing and maintaining a complex web of interfaces for the exchange and conversion of data between applications.

To cope with those difficulties, the Bank decided to develop a business intelligence (BI) architecture, to be used as reference in all future IS/IT developments in the statistical area and capable of promoting efficient data analysis. The adopted BI framework aimed to contribute to the construction of a coherent statistical information system for the *Banco de Portugal* (*i.e.*, one that is not just the result of a simple juxtaposition of the ISs in each domain), by creating a framework for the development of the systems focused on the goal of integration. Such framework was built upon three pillars: (i) a reference data management centre, (ii) a data warehouse (DW) and (iii) a common IT platform. The centralised reference database, developed on the basis of the already existing *Reference Information Sharing System* (or SPAI, using the Portuguese acronym), provides the connecting elements of statistical data from different sources and, ultimately, is the main guarantee of the possibility of integrating the information, enabling cross-linking information from different sources and systems; the DW guarantees a central access point to every statistical data, independently of the input source or the production process; a common technological infrastructure across multiple information systems makes it easier to integrate and reuse components and promotes data access efficiency and transparency to final users.

A DW strategy typically falls in one of the following cases:

– *Top down* approach, whereby a central DW is first constructed, from which are then extracted specific data marts for each business process. It allows the integration of multiple information domains, achieved through a global development effort. However, its intrinsic difficulty, together with the time and cost required to obtain the first results, explain its high failure rate.

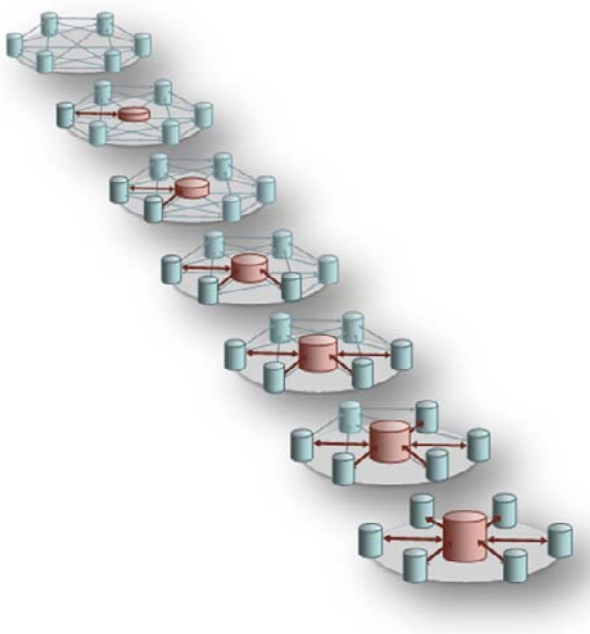
– *Bottom up* approach, whereby several data marts are first developed, on the basis of which the global DW is incrementally built. This allows for reducing the risk of the projects and speeds up the achievement of the first results. The problem with this approach is that it often results in new information silos, when successive isolated projects fail to promote the integration of data from different sources.

The Bank opted for the second approach, on the grounds that in our case there is an intermediate area – the so-called "working data store" –, where production processes take place, accessing data from different domains already stored in the DW – unlike the more common situations, in which data flows from source systems only in one direction and the construction of the DW has no upstream impacts. In fact, the additional complexity arising from this dual dependency, in which information providing processes are at the same time clients of the DW, suggests following an incremental approach.

Figure 2. illustrates a schematic representation of the evolution from the method of direct integration of various production systems to the final model, in which the integration will be done through a central repository of information coherent and consistent.

Phases of the incremental approach

Figure 2



As the number of embedded systems increases, the number of data streams becomes smaller. In this representation, the first attempts at implementing the architecture do not result immediately in a significant reduction of complexity, since all the other systems are still dependent on the current interfaces. Clearly, this process has the potential to last for too long.

Conversely, at later stages of the process each new integration project will allow a further reduction of the existing interfaces, which will be replaced by access to the central repository. This implies some intervention aiming to ensure compatibility with the systems already integrated (which is the cost of the incremental strategy).

One thing to take into account as regards the development of this paradigm is that the projects beyond its scope and objectives must be subordinated to a broader program and abide by a common reference, so that the DW emerges naturally from a set of shared but coordinated efforts.

Concluding remarks

The strategy followed by the Statistics Department to develop a BI architecture was based on the following principles:

1. Incremental development of the individual projects as a way to reduce risk and achieve rapid results.
2. Project monitoring by a joint IT Department/Statistics Department committee, to determine whether the objectives of the project are being successfully met and to keep the BI architecture updated.
3. Shared responsibilities with the IT Department, thus benefiting from its specific technical knowledge and, by the same token, securing the degree of autonomy necessary to meet the changing requirements and business rules, while guarantying reliability, performance and safety to the structural components of the solution.
4. Post-implementation support on the part of the IT Department, to accommodate new requests without compromising the performance of the system, and to promote proper use of tools.

On the basis of our experience we were able to identify and to deal with a number of key success factors, *inter alia*:

1. Rethinking the governance model adopted in 2008, with a view to rectifying a number of insufficiencies that were hampering its effectiveness. A new BI management model is being implemented, involving both the Statistics Department and the IT Department.
2. Taking stock of the information available in the different areas of the Statistics Department (creating data catalogues). In this context, the meta-information model is of paramount importance.
3. Assessing the capacity of the existing infrastructure and whether such infrastructure is appropriate for the BI project that was originally designed.
4. Ensuring a greater operational flexibility of the relational databases and cubes in the mix of solutions to adopt.

5. Overcoming the limitations of the current analytical systems, as regards its capacity to produce new metrics and/or indicators.

6. Training "power users", capable of providing support to other users.

In spite of the difficulties and complexity of these challenges the Statistics Department was able to create the minimum necessary conditions for data integration. Our micro-databases are now based on common reference information, allowing for new statistical products to be launched and new processes to be implemented, with clear advantages to the users. However, if it is safe to say that we are starting to reap the benefits arising from having a higher level of data integration, there are still many issues pending, which require new approaches and extensive research and development efforts.

References

D'Aguiar, L., de Almeida, A., & Casimiro, P. (2011) "Promoting enhanced responsiveness to users' data needs: the experience of the *Banco de Portugal* in exploring the statistical potential of micro-databases". *Proceedings of the 58th ISI Session*. Dublin, August 2011.

Lima, F. & D'Aguiar, L. (2009) "Credit risk transfer – Dealing with the information gap". *IFC Bulletin,* No. 33, August 2010.

Martins, C. (2008) *et al*. "Arquitectura de BI – Sistema de Informação Estatística", *Banco de Portugal*, June 2008 (not available in English)

Martins, C. & Aguiar, M. (2011). "Adding business intelligence to statistical systems – The experience of Banco de Portugal". *Eurostat Conference on "New Techniques and Technologies for Statistics"*. Brussels, February 2011.