



**Economic and Social  
Council**

Distr.  
GENERAL

CES/AC.71/2001/3  
22 November 2000

ENGLISH ONLY

STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN  
COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/Eurostat Meeting on the Management of Statistical Information Technology  
(Geneva, Switzerland, 14-16 February 2001)

Topic (i): The impact of data warehousing on the management of statistical offices

**JUST IN TIME PROCESSING AS ONE OF THE REQUIREMENTS FOR INPUT DATA  
WAREHOUSES**

Submitted by Statistics Netherlands <sup>1</sup>

**INVITED PAPER**

**I. INTRODUCTION**

1. During recent years, Statistics Netherlands has been experimenting with data warehousing techniques. Amongst other things, we found that there is a need for organisational change if one is determined to apply data warehouse techniques to production. In this paper we will explore the type and extent of organisational change that is needed. For the technical part of this paper we adhere to the definitions of Kimball.

2. The greater part of this paper will concern the changes that appear to be needed in the foreseeable future, since the use of data warehouses is still in its infancy at Statistics Netherlands. However, we believe that lessons can be learned and shared from our experience.

---

<sup>1</sup> Prepared by M.H.J. Vucsan.

## **II. WHAT IS A DATA WAREHOUSE AND WHY IS IT TIME CRITICAL?**

### **II.1 The nature of a warehouse**

3. A data warehouse is more than a massive amount of data in a storage module or filing cabinet. A data warehouse is a massive amount of organised data. This means that the data are grouped by subjects and that metadata are present. It also means that the data are stored in such a way that query is simple and easy and can be performed with standard commercial tools. Such requirements are the reason why data are stored in data marts, subject oriented modules. The data marts make use of dimensions that contain the possible textual headers (i.e. the meaning of the figures!) For every subject there is a specific template and process as to how the data are entered. There is little room for variation. Typically, data are added to a data warehouse, and alterations are not very common.

### **II.2 Cumulating events is time-critical**

4. As opposed to a typical online transaction system, such as a flight reservation system, a data warehouse is not meant to continuously reflect the current state of affairs. In a data warehouse facts are accumulated over time. Unless driven by lack of storage space, it is unusual to delete older data from a warehouse. Updating existing data to reflect the more recent situation or viewpoint is also not very common. A data warehouse can thus be seen as a time series. Selections from a data warehouse will nearly always incorporate a mention of time. This cumulation of events is very much the added value of a data warehouse.

5. Adding facts means that the facts to be added must be complete. Updating facts is not very practical and sometimes dangerous. This means that if we want to add a person to a city at an address which was recently built, we have to have the specifics of the address before he or she can be entered into the data warehouse.

6. Entering data into a warehouse means that all the dimensions must be up-to-date. For the primary dimensions which are derived from the incoming data this will be an automatic process, but for the secondary dimensions produced by other departments or other processes this is problematic. The existence of secondary dimensions is due to the fact that not all the knowledge needed is included in the data. Think about postal codes and coordinates of addresses. Although we receive the address in the data, we need a file with the extra data about these addresses (coordinates, postal code, etc.). Another department, with its own timetable, produces this file.

7. Loading a data warehouse means that lots of data come together simultaneously from different departments or taskforces in the agency. Therefore the preceding cleaning processes are time critical.

### **II.3 The input and cleaning process**

8. The first important problem is timing. Although files are being exchanged and distributed in the bureau this is not always subjected to rigid production planning. So when loading a data warehouse, you sometimes would have to wait for files that are needed for certain data warehouse dimensions.

9. A second problem is the way in which the input is processed. Most statistical agencies receive quite a lot of data in very different formats. We have our own field research but also accept files or complete registers from other institutions. Up to now the common wisdom has dictated that this data is put through batchwise cleaning processes. It is here that the integration problems begin. After cleaning, data about the same subject from different sources may produce different results that may be caused by different cleaning algorithms and/or definitions. The main cause of this problem is also called the stovepipe architecture of statistical agencies.

10. In this architecture the accepted way of processing is the refining of files. Again and again the files are processed and a new generation of files is produced. As a rule no information is kept in the data about the nature and extent of the imputation done. Therefore, there is no easily accessible information about the error margin of the data or the percentage of imputation done.

11. Data warehousing requires another approach. Not only is a more centralised way of input processing needed, it is also fruitful to separate the basic cleaning (format checks, etc.) and the cleaning of a statistical nature (plausibility, coherence with previous years, etc.). The basic cleaning can be done in specialised production systems either automatically or manually with little relation to content. This basic cleaning and input gathering has a just in time characteristic.

#### **II.4 Thou shall not modify...**

12. When data is cumulated into a data warehouse, the problem is that a fair amount of the data may be incorrect. Traditionally, there is some cleaning process in place in order to produce data, which will deliver the correct statistical answers (there may be artefacts in there!)

13. In a data warehouse this practice may lead to strange results. For instance you cannot “fix” the number of cars present in Amsterdam by just assigning a car to a number of people that have none. In a data warehouse the statistical analyst will be able to see that these are people without a drivers licence or perhaps with not enough money to buy a car! For this to work, you will need a data warehouse dimension that will indicate which facts are observed and which facts are the result of cleaning activities. You will have to do the cleaning in the data warehouse by adding corrective facts in a meaningful way.

14. With a dimension that indicates the source of the data, we will be able to give an indication of the margin of error in the result figures. This is because we can always separate the observations and the added corrective data. The important idea is that a complete history of the corrections be kept.

### **III. ORGANISATIONAL IMPACT**

#### **III.1 Integration of the input process**

15. Up to now the priorities for producing files have been quality and then timing as long as they were delivered within the production interval of the statistic stovepipe. However, if you start integrating the input phase it becomes clear that timing is everything and that quality is to be achieved in a continuous fashion and not at discrete intervals. This means that the idea of separate groups processing specialised data can no longer be sustained and other structures are needed.

16. Some time ago Statistics Netherlands started to concentrate on input activities and abandoned the

idea of stovepipe architecture. Initially, it was conceived that one input division would be the best solution, but later it became clear that the two main fields of interest, economic and social statistics, should have their own input department.

17. With the organisational concentration of the input activities, better integration and timing are possible. Although the data are not really stored in a data warehouse at the moment, it is to be expected that future developments will soon lead to this. Already a separation has been made between technical cleaning and statistical cleaning of the data.

### **III.2 Moving away from discrete processes**

18. The best way to cope with massive amounts of data is to not let the processing pile up. Therefore there is a need to load the data warehouse at the same rate at which the data is coming in.

19. When loading data continuously in a data warehouse, the whole idea of processing data in large batches has to be abandoned. In this context, large batches mean that the main focus is to achieve some point in time where “all of 1999 or so” is processed. Most of the time, the interval is equal to the production interval of the statistic that was the original producer of the file. The whole organisation was focused on producing preliminary and final data. A data warehouse needs all the related data at the same time. Therefore it is imperative to process all the input as one integrated process.

20. With input integration it is probably preferable to move from a system with preliminary and final figures to a system with a continuous margin of error. With a continuous margin of error the accuracy of a statistic would increase with time. Coarse figures would be available very soon, at least to the analysts in the agency itself. For the general public it would represent the possibility of making an explicit decision between timeliness and quality. It is the logical result of ongoing pressure to produce more figures at shorter notice.

### **III.3 Administration as success factor**

21. Data warehouse administration is about content, it is not an IT job. In our view the data warehouse administrator is part of an administration hierarchy that closely follows the way the statistical divisions are organised with respect to content. A content related cluster of data marts should have an administrator who “lives” close to the group of analysts that use the warehouse. On the top level of the statistical division there should be a chief administrator who is also responsible for the security of the data warehouse.

22. In statistics a large majority of the variables are selection variables. This means that they will be stored in dimensions. For data warehousing to succeed it is imperative to have an elaborate administration of these dimensions. The dimension administration office could very well be a part of the data warehouse administrators hierarchy.

23. At the moment, at Statistics Netherlands, we have not yet implemented such a structure. However, this does not mean we are not involved in some kind of central dimension administration. At the moment a central classification server is active where all the major classifications are to be stored for general use. Classifications are a precursor to dimensions in the sense that they can be seen as elementary building blocks.

#### IV. CONCLUSION

24. Although not yet implemented at Statistics Netherlands, it becomes clear that, for deployment of data warehousing techniques in the input stage of processing and further down the line, there are some guidelines that can be formulated:

- ◆ there should be centralized processing of input and centralised planning of the input and technical cleaning process to eliminate timing problems. Centralised in this context means centralised to the level of statistical division or main area of interest;
- ◆ processing structure and therefore process control should be geared towards continuous processing of the input streams and not towards statistical production interval. Publications should carry a margin of error which may narrow as time progresses;
- ◆ it is important to have an elaborate data warehouse administration structure with a hierarchy up to agency level in order to preserve security and to foster statistical integration. This includes dimension administration.