

SEMINAIRE

E+; 3=! C

SEMINAR

STATISTICAL COMMISSION AND
ECONOMIC COMMISSION FOR EUROPEDistr.
GENERALCONFERENCE OF EUROPEAN
STATISTICIANSCES/SEM.43/4 (Summary)
14 January 2000

Original: ENGLISH

Seminar on integrated statistical information
systems and related matters (ISIS 2000)

(Riga, Latvia, 29-31 May 2000)

Topic I: Data warehousing and the development and use
of statistical databases in a network environment

**EVALUATION OF A WAREHOUSE SOLUTION FOR CORPORATE DATA MANAGEMENT
IN A NATIONAL STATISTICAL OFFICE**

Invited paper

Submitted by the Swiss Federal Statistical Office¹

SUMMARY

I. PRINCIPLES

1. According to Sundgren, the corporate warehouse will form the backbone in the architecture of a national statistical office. It is the repository for all final microdata (final observation registers) and all valid macrodata (including indicators) together with the corresponding metadata. The warehouse has to provide tools for carrying out, as an integral part of its functions, the majority of the transformation processes from micro- to macrodata. In terms of metadata, the warehouse has to cover and manage official terms of universes/populations, objects/variables and all master classifications, the description of transformation processes as well as the inventory of sources, activities and statistical products. The warehouse should be the source for an on-line database with (fee-based datamart) access for external users (one form of access being Internet) with adequate search and browser tools including data export.

II. EXCLUSION

2. The warehouse will not cover data processing activities prior to final observation registers with the exception of interfaces for data and metadata exchange between the central repository and the production environment. The management of registers of units (such as business registers) and of geocoded microdata or grid-

¹ Prepared by Georges Fleuti.

defined aggregates is also excluded. Furthermore, we do not address the various ways of creating output for dissemination from the warehouse.

III. FUNCTIONS

3. As a backbone, the warehouse has to cover the major functions needed in the statistical process in order to minimise all routine work. Most of the functions have to be driven by metadata. The loading of raw data from various sources into a relational database for microdata has to be as simple as possible and minimise all routine work. This microdata has then to be transformed by automatic aggregation into macrodata-cubes in a multidimensional database. Indicator cubes combining elements from different macrodata cubes are the third level of the warehouse. One major function is the access to all data - micro and macro - for ad-hoc queries. Microdata has to be accessible online to internal users for the purpose of the selection of subsets or to build ad-hoc macrodata.

IV. REQUIREMENTS

4. In terms of languages, the system has to offer multiple languages both for user guidance and metadata. High-level performance and user-friendly interfaces are unavoidable. The system has to offer the possibility of implementation on a variety of IT-platforms as well as to ensure a good performance for all functions involved.

V. POLICY

5. For the successful and efficient functioning of a warehouse in a statistical office, it is essential that a strong policy be set up and implemented to ensure that no output characterised as official statistics disseminated in any form contains any figure or concept related to figures that are not present in the macro part of the warehouse. The way of structuring and defining the macrodata to be included in the warehouse will have to follow certain rules and criteria and cannot be decided unilaterally; it includes the definition of corresponding metadata. The access to the warehouse is strictly limited to in-house users.

VI. BENCHMARK

6. A benchmark is a good opportunity to prove whether the offered function really exists and how different database systems perform in all the relevant steps in the whole warehouse process for a given platform.

7. The core elements for the benchmark can be subdivided in 5 main parts:

- Architecture,
- Database system (microdata, macrodata, metadata),
- Functions for selection of microdata and macrodata,
- Same metadata for microdata and macrodata,
- Transformation of microdata into macrodata by aggregation.

8. The benchmark organised by the Swiss Federal Statistical Office in 1999 within the warehouse project produced surprising results. As the common platform for the tests, the Compaq-Benchmark-Centre in Valbonne (France) provided a 4 processor 64 bit Alpha-Server with 1 Terabyte of disk space in RAID-5 technology. The operating system used was Tru64-UNIX. The three partners - ORACLE, SAS and MSI agreed to compete and obtained 4 full days to fulfil the requested process.

9. The initial data provided on flat ASCII files were taken from the population census of 1990 from foreign trade and data from a sample survey on rents of dwellings with the common structure of the municipalities. Data was then replicated by special

algorithm by 50 years, to obtain over 2.2 billion records to be stored on 106 Gigabytes of disk space.

10. The comparable results focused on space needed, run time for batch jobs and response time for ad hoc queries. The differences between the three systems were considerable, ORACLE and SAS being factors (up to 10) slower than WIDAS of MSI. The results are specific for the warehouse environment of a statistical office, which is not oriented towards transactional business but rather to queries whose standardisation is limited.