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

Incremental Implementation of the Data Warehouse

In the Central Statistical Office of Poland

Submitted by CSO, Poland

CONTRIBUTED PAPER

I.   Introduction

1.   The value of qualitative as well as quantities information on any subject related to macro and micro-economy is increasing everyday in the neighborhood of statistics. As a part of human activities related to exchange of information on current situation of national economy and society as a whole statistics has many obligations to fulfill.

2.   To assure high quality of end-user service it is advisable to design and implement a statistical data warehouse. The Data Warehouse as an assumption should fulfill some important goals like:
♦   possibility of easy statistical survey analysis based on distributed sources of data,
♦   support for the process of calculation, aggregation, and presentation of statistical analysis,
♦   integration of the results of statistical surveys,
♦   support and increased efficiency of data dissemination and data access for end-users,
♦   simplified administration of metadata and databases.

I.1   Central Statistical Office in the External World

3.   The overall schema of the CSO communication with external users is presented on the following picture [1]:
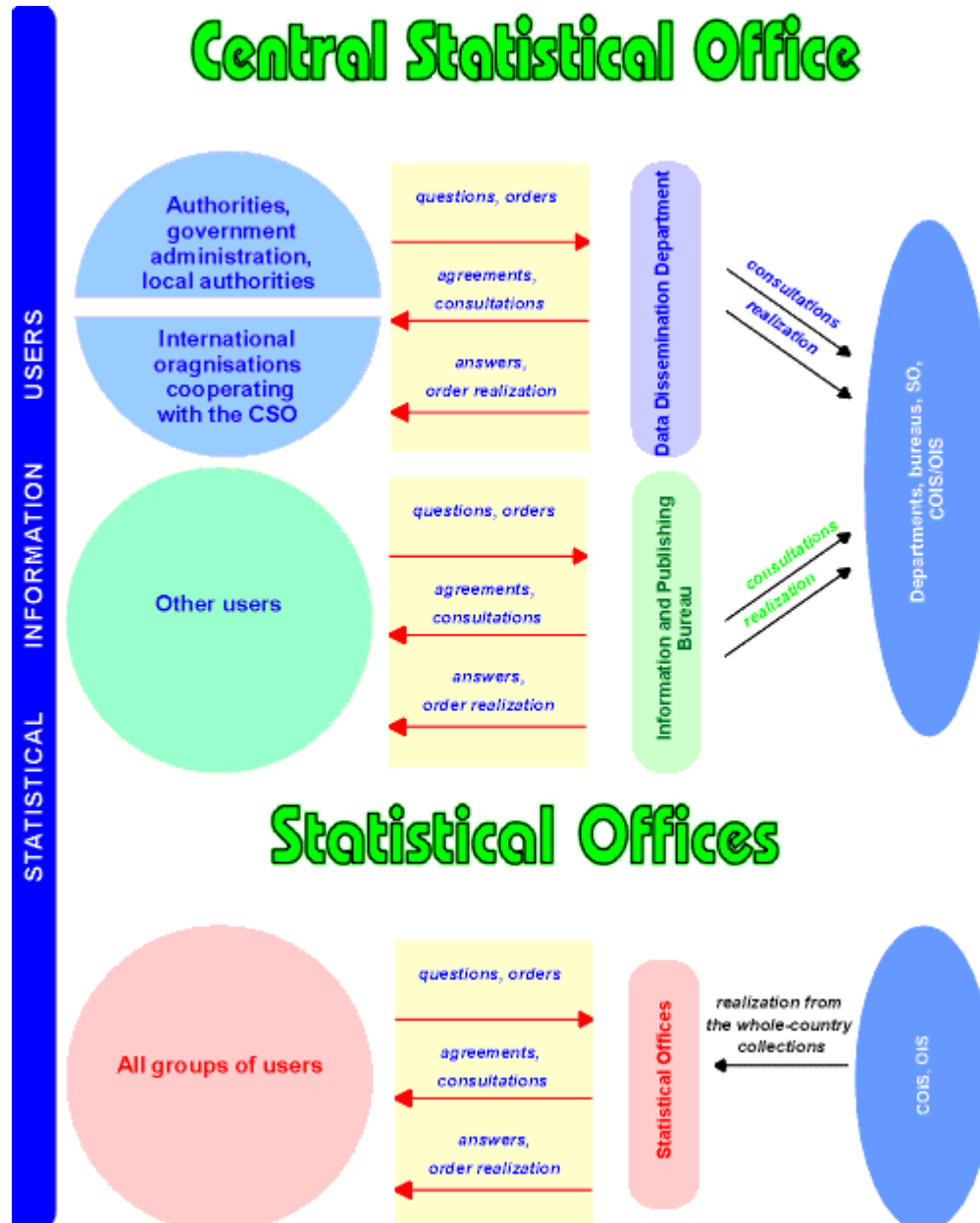
4.   There are many ways, to organize the efficient access to the results of statistical surveys, selected data aggregates, and publications, including data distributed electronically, but the most efficient is to install powerful WWW servers with an access to the data warehouse with stored information, and with possibility to use OLAP tools for those who are experienced end-users.

---

[1] Source: http://www.stat.gov.pl/

I.2 Existing infrastructure and development plans

5. We have to start our preparation for a data warehouse implementation from the current limited resources and infrastructure, but we have to take into account options and possibilities of further investment and development work. At the time of writing the paper the IT infrastructure at the statistical offices of Poland was as follows [2]:

Computer	Operating System	Number
Servers		
HP 9000/K400	HP-UX 10.20	1
HP 9000/800	HP-UX 10.20	18
HP 9000/700	HP-UX 10.20	3
HP LC P100	Win NT 4.0	19
Compaq P III 800	Win NT 4.0	22
Pentium 450/500	Win NT 4.0	13
Compaq P200	Win NT 4.0	64
Pentium III 733	SCO Unix 5.0.6	50
Pentium II 350, HP LC P100	SCO Unix 5.0.2, 5.0.4, 4.2,3.2	158
Workstations		
Celeron 600	Win 98	1608
Pentium III 450 – 600	Win 98, 95, Win NT Workstation	160
Pentium II 350	Win 95	139
Pentium 75 – 166	Win 95, 3.11	1443
Terminals		
486 DX, SX	Win 3.11, (Unix)	819
386 DX, SX	DOS, Win 3.11, (Unix)	825

6. To prepare our infrastructure for successful realization of the National Population and Housing Census 2002 (combined with the Agriculture Census) we have to invest in many areas, but from the IT point of view the following are the most important [3]:

♦ restructuring of the CSO Wide Area Network,
♦ extension of Local Area Networks in approximately 20 localizations,
♦ Server, RDBMS, and processing tools for the Central NC Data Warehouse,
♦ Servers for Statistical Offices for storage, editing, and processing of the NC data,
♦ Scanners for automation of the process of digitalization of data from the NC forms,
♦ Hardware and software to secure environment for data storage and processing,
♦ Additional equipment for desktop publishing and Statistical Publishing Establishment.

---

[2] Based on the document: „Assumptions for data storage, data processing, and data dissemination for the National Census 2002", Team work, CSO, Warsaw, December 2000.
[3] Excerpt from the budgetary proposal for the year 2001, CSO, Warsaw, November 2000

7. Future investments to be covered both from budgetary funds and EU projects:

Investment Title	Predicted Scope and Functionality
Server for WAN administration	To fulfill requirement of efficient work of the Corporate WAN
Data dissemination server	Efficient access to the results of surveys with special consideration for the NC data
FIREWALL Server	Increased protection of WAN and statistical data against intruders
TERYT Server	GIS background information for the localization layer
BDR Server	Regional Data Storage and Dissemination Servers
World Wide Web Server of the CSO	Dissemination of all data with the use of Internet
Library Server	Central Statistical Library Information System
Polish Server for EUROFARM data	Central Database Localization (may be as a Data Mart)
Departmental Database Servers (Data Marts)	National Accounts, Demography, Tourism, Labour Statistics, Indicators
Modernization of the Central Servers and the Application Development Tools	Scalable development of the Data Warehouse environment through introducing of new technologies and new data marts into global Data Warehouse for Public Statistics

II. Background of requirements for a data warehouse design and implementation

8. At the end of 1995 during the elaboration of national accounts at the Statistical Office in one of the districts a team working on the subject realized that collecting information from many sources, which are distributed across many platforms and different software is a daunting task. All of these sources (internal and external) were in different formats what required transformation and putting them in order, what gives a possibility to compare the data and figures resulting from calculations. The team suggested to start appropriate ordering and cleansing of data in national statistics, taking into account introducing of standard data description (metadata), putting in order used naming conventions in the area of nomenclatures and classifications used in the area of statistics. There have been suggestions to extend the standardization process for database dictionaries and structure of tables used to store atomic-level data, aggregate data, analysis results and publication materials and documents as well.

III. First organizational steps for future data warehouse development

9.   In September 1996 the Chairman of the Central Statistical Office of Poland decided to create a team that will be responsible for coordination of all the related tasks, which will lead to complex statistical database. According to assumptions the group of experienced statisticians (including middle management), IT specialists, and responsible staff of the Central Statistical Computing Center have been engaged in the following tasks:

♦ Collection of information about statistical surveys under consideration,
♦ Definitions of atomic-level data collected with the use of statistical forms,
♦ Description of rules and methods of data aggregation,
♦ Specification of data analysis processes fulfilled during elaboration of statistical surveys,
♦ Preparation of lists of end-users groups, who utilize the results of analysis,
♦ Elaboration of policy for statistical data and IT infrastructure security,
♦ Compilation of rules for links between different sub-systems and their design,
♦ Overall framework of the metadata database and main statistical database development
♦ Introducing of rules for creation of the thesaurus of statistical terms,
♦ Listing of main elements of planed databases on content of the database dictionary
♦ Other tasks and steps related to subject.

III.1 Concept of dedicated data base for processing statistical surveys

10.   Because during the time of first elaboration, data warehouse terminology was not yet introduced into everyday practice in statistics, the name of "Common Database for Statistical Surveys" (CDSS) have been used to describe the data warehouse for public statistics. All involved members of the team realized very soon that the set of tasks is very complex, and that each of the tasks is a difficult problem on its own. In the mean time knowledge about theory and practice of the data warehouse was growing among members of the team, and other statisticians as well. Moreover, in the written documents the term CDSS was used in exchange with the name "Data Warehouse for the Public Statistics" (DWPS), and we are using both terms as synonyms.

11.   Consecutive meetings and discussions between members of the team and much wider group of statisticians lead to compilation of final document that was presented in November 1998. The document contains overall framework of work that should lead to creation of the DWPS. Description of time frame and hardware and software tools required to the DWPS implementation have not been presented, but logical background with specification of subsequent tasks that should follow the study have been described quite precisely.

12.   A number of teams linked to the particular problems were involved in detailed analysis and work on the concept of the DWPS. Among other tasks, the teams have been trying to establish the

design of the first tables, that should be used as conformed dimensions in the future data marts and that will constitute the basis for metadata development. Another sub-set of metadata should be used for description of methodology of statistical surveys. At the same time the creation of full thesaurus of data from its atomic level up to different levels of aggregation has been foreseen and it is implemented now.

13.   During the discussions and work on the separate elements of the concept there appeared new elements of complication during consecutive steps in the process of analysis. When one problem was solved sometimes few others even more difficult appeared. For example, in every year practice in statistical surveys in Poland both methodologies as well as survey realization usually were changed dynamically, very often in quite a big amount of surveys. Moreover, design of forms for data collection was also very often changed from year to year from survey to survey. This was leading to extreme overloading of IT staff, especially in units responsible for country level surveys.

14.   The new approach was proposed, to create thesaurus of items (facts) that should be collected in the fact tables together with description of related conformed dimensions. It has been suggested to automate the form generation, to simplify its design, to reduce related workload and at the some time, to prepare a background for future generation of electronic forms for data collection with the use of Internet.

IV.   Framework for the CSO Statistical Information System

IV.1. Cooperation in the frame of EU

15.   Harmonization of the Polish Statistical Information System with the EU requirements is being implemented since 1995. Important elements of those tasks have been financed from EU funds. Also in the current project PHARE 2001 there are elements supporting database development which are in line with the CSO framework of the data warehouse design and implementation.

16.   Sub-project III of the PHARE 2001 projects assumes development of the metadata database as the key component of the statistical information system and the CSO data warehouse including the following elements:
♦ Methodological description of the statistical data,
♦ Statistical terms thesaurus and dictionary,
♦ Localization of the terms on the statistical survey forms,
♦ Links to variables and columns in the source data tables,
♦ Definitions of terms used in statistical surveys,
♦ Survey methodology,
♦ References to the Polish/EU legal base,
♦ Description of methods of access to the statistical data,
♦ Descriptions of the statistical data files.

IV.2. *Initialisation of data loading into the DWPS*

17.   We have decided to start loading of data to our Data Warehouse for the Public Statistics beginning with the National Population and Housing Census 2002 data, and National Agricultural Census 2002 data. Afterwards we have decided to load data from our standard statistical production systems. Firstly we will concentrate on the following systems [4]:

♦ C-01 – Prices of products, services, and works,
♦ Demographical systems (births, mortality, migrations, divorces),
♦ DG-1 – Monthly information about commercial activities,
♦ F-01 – Income, costs, and financial results of enterprises,
♦ F-01/b – Financial statements of banks,
♦ F-01/m – Financial statements of financial institutions,
♦ F-01/s – Financial statements for universities and high schools,
♦ F-02 – Statistical financial statement,
♦ F-03 – Statement of used assets and carried out investment,
♦ I-01 – Statement of investment activities,
♦ RB-xx – Yearly statements of budgetary expenses and income,
♦ SP – Yearly statement of enterprises activities,
♦ SP-3 – Yearly statement of activities for commercial firms and institutions,
♦ Z-01 –Statement of employees,
♦ Z-03 – Statement of salaries and employment,
♦ Z-06 – Statement of employment, salaries, and working time,

18.   Currently our Central Statistical Computing Centre is concentrating on preparation of overall specification for the future data warehouse environment, writing down CSO wishes and demands regarding RDBMS system to be used. Those specifications will be distributed to local (Polish) dealers selling RDBMS and the data warehouse solutions and after collection of their offers and proposed solutions we will decide about proper investment in the hardware and the software platform for our future data warehouse.

V.   Basic structure of the DWPS

19.   We are planning to divide our data warehouse into several sub-systems, this means:
♦ data cleansing, transformation, and loading system,
♦ metadata system,
♦ database system for data warehouse (including data marts),
♦ data dissemination system, and in this:
  o OLAP system,
  o data mining system,
  o Web access system,
  o predefined reporting system.

---

[4] „Introductory conditions for design and implementation of statistical data warehouse in the CSO", CSO,

Warsaw, 2001

20.   Overall schema of the data flow between different sub-systems can be presented as follow:

| Data loading | Data warehouse | Dissemination |
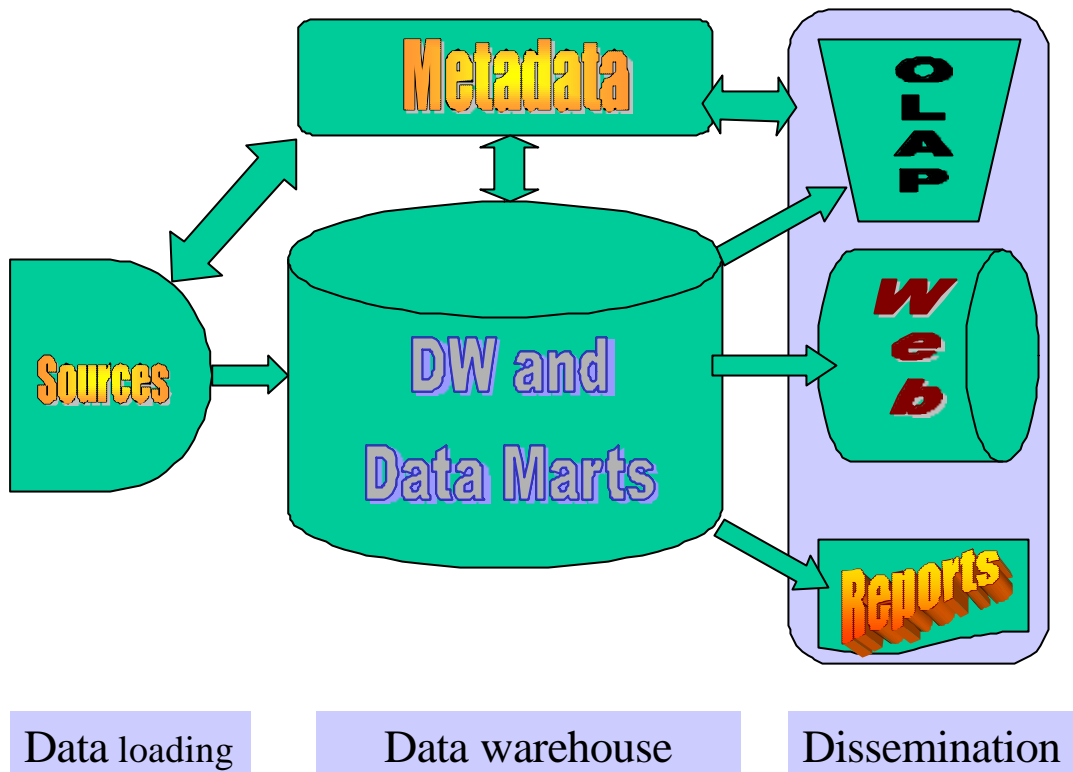
VI.   Conclusions

21.   Task of data warehouse implementation in the CSO of Poland will be realized in the years to come. We are going to create data marts, one by one, and link them into the overall DWPS structure. Our aim is to prepare scalable environment both in the hardware and the software and tools area, so we can steadily increase the number of end-users and efficiency of statistical data dissemination processes.