Business Intelligence Tools for Statistical Data Analysis

Savoska Snezana, Manevska Violeta
Faculty for Administration and Management of Information systems, University "St.Kliment Ohridski" Bitola, Partizanska bb, 7000 Bitola
snezana.savovska@uklo.edu.mk, violeta.manevka@uklo.edu.mk

Abstract. *Dealing with the up flow of information is one of the most significant tasks of the information systems nowadays, especially when they are official information from state institutions. Getting information from all state sectors is a very complicated task, dedicated to the State Statistical Office (SSO). This task cannot function without using highly sophisticated Business Intelligence (BI) tool for Data warehousing and information visualization. The objective of this paper is to elaborate the study on utilization of BI tools for complex statistical calculations in the SSO in the Republic of Macedonia. This tool has been implemented in SAS software modules, sophisticated statistical software for data extracting, cleansing, transforming and loading in Data Warehouse (DW).*

Keywords. Business intelligence, Data Warehouse, Statistical data analysis.

1. Introduction

The SSO DW is the most professional statistical usage of DW for BI in the Republic of Macedonia. It is installed on the SAS DW software environment with precisely defined procedures for extraction, cleansing, transformation and data loading. The physical model of SSO DW is shown on Figure 1.

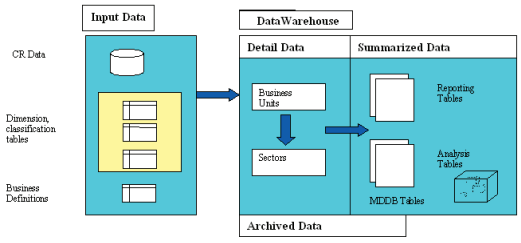

Figure 1: Physical model of State Statistical office DW

Data which is an object of data processing in the DW is taken from different administrative sources, areas from where data is obtained with statistical researches made in SSO and from Statistical Registry. They are extracted from many sources, prepared with manual or automated software procedures, transformed in specific code pages and loaded in temporary DW files.

2. The Data Warehouse content

The administrative data sources are: The State central registry (the final account report of subjects), The office of public finance (the taxes-information database), ministry of finance (public revenue – the treasury accounts), National bank (monetary statistics), Sanitary insurance fund (contribution of sanitary insurance) as Retirement and invalid insurance fund (contribution for retirement and invalid insurance).

The statistical researches are made in the areas of: Inside trade market, foreign trade market, industry, building and construction sector, tourism, agriculture, poll for household spending, prize statistic and poll for labour.

The data which are loaded in DW are taken from different data sources, working in different operating systems and platforms. First, they are processed with Microsoft Visual Studio.NET (Visual basic or ACCESS) or other preparation procedures. They use SAS Data integration studio for data import (Figure 2), selection and transformation. It is a very powerful tool, which can provide integration of diverse databases laid in different data servers or from external sources. Also, it can provide creating of multidimensional cubes, modelling of software processes and creating calculating methods for summarized (aggregated) tables and DW. These procedures are used for finding irregular or missing data, transformation of these data, conversion of Macedonian code page in ACSII cods, prepared

in format suitable for first run processing and loading in DW. There are many control procedures, specific for data from different sources and also procedures for data transformation and loading in DW. Some of these procedures are manually made, because of its specifics and it can't be automated.

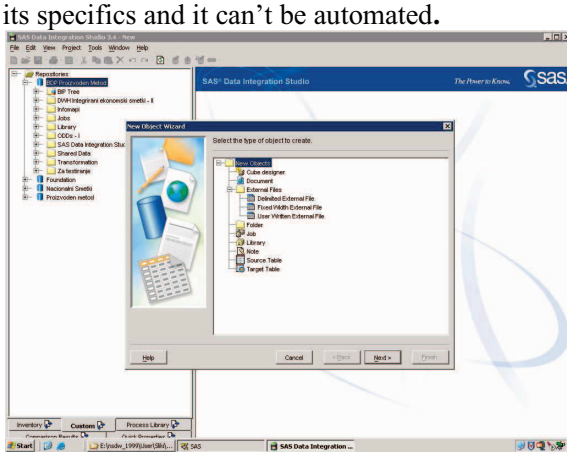

Figure 2 – Screenshot of SAS Data Integration studio

3. Data preparation for Data Warehouse

The data loaded in the first phase are checked with SAS DW coded program. It possesses business logic that includes data transformations, depending on the needs of the statistical processing, summarizing and other data transformations. These procedures are called Firs run processes.

.

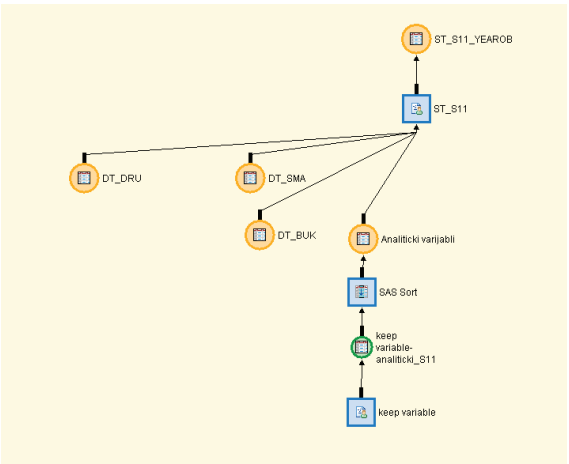

Figure 3: Modelled software business processes with defined program code

These defined processes and procedures made on the data in the DW perform gaining of the desired outputs. They are defined through SAS process designer, visual tools that depict all defined and needed processes. Depending on the type of business processes, they are mapped in a diverse way – objects with different shape, colour and functionality. We can see detailed depiction of software business process and what it is a doing with the data. Also, we can see the original program code for each object with double click on item. The visual representation of one of the modelled business processes is shown on Figure 3.

4. Metadata of the Data Warehouse

Because the metadata are information about data, can be defined in many different way. There is couple of metadata types: input data metadata for DW, metadata of data cleansing and transforming, semi-resulting data and target data metadata with its aggregated tables stored in DW. Some of the metadata descriptions are stored in the documentation and code files of DW, and the transformation of data and the used method for this transformation in the semi-results or final outputs are placed in an Excel file. This file contains the business logic for calculating the needed statistical ratio and the other desired statistical information (Figure 4).

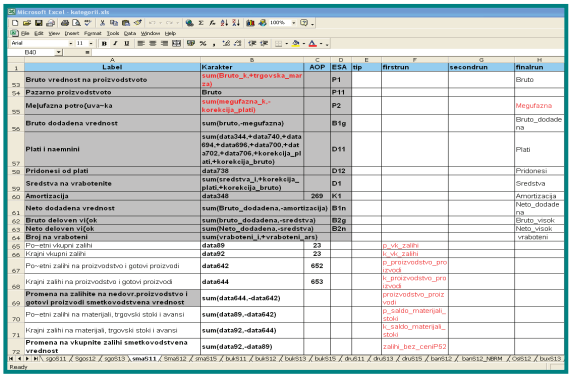

Figure 4: Excel file which take a role of metadata defined file

With all these transformations, we calculate the new variables, called category for calculation of GDP. The final outputs are settled in many different summary tables, grouped depends on institutional sector. This data are disseminated as public data and placed in the SSO web site.

For updating data of the State central registry -the final account report of subjects in the DW, data is merging with dimensional tables of each subject of DW being made. The input data from

the State central registry, obtained in ACCESS tables and saved in the temporary folders, is cleansed, recoded and transformed. Then, the data is loaded with SAS in a DW Integration studio. Cyrillic data is recoded, merged with identified data in the subject's data of the State central registry and some additional variables are added in the dimensional tables. This additional data is prepared for the next phase of data preparation. These dimensional tables have the territorial classification, valid in the country, the institutional sector which possesses this data, National classification of activities (NCA) and other.

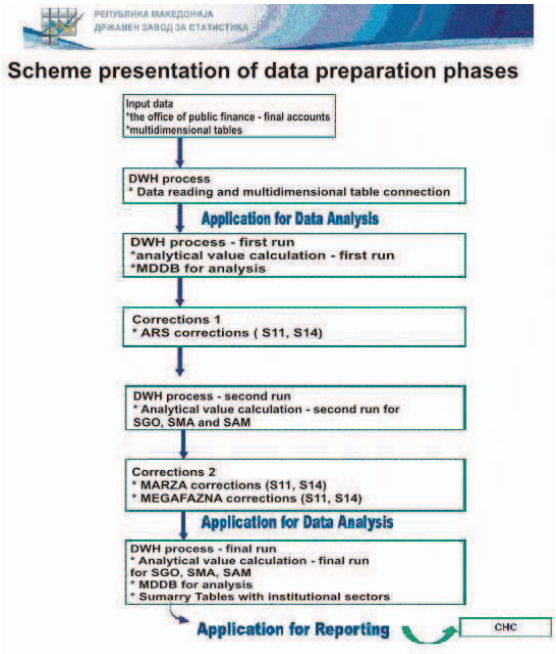

Figure 5: Schematic presentation of data preparations phases

A very important subject who gives additional information for business subjects and institutional sectors is the National accounts sector (NAS) has the responsibility for all category data calculation, keeping the formula up to date and data control. This data is defined as metadata for the transformation steps and is saved in "Kategorii" Excel file. This is the most important file for making all data calculations and must be updated and also, the oldest "Kategorii" file must be saved for historical analysis. This file is the base for the intermediate file creation and for the next calculation steps. This phase is called First run and the result is aggregated tables with all the data, but unlashed of grey economy (illegal market). The second run programs enrich these calculations with mythological solutions for embedding some influence of the grey economy. The data from these two phases is compared and ultimately made some corrections. Then, the Final run phase can be started. With this calculation phase, the ESA codes for sector accounts are prepared. This phase is called also ESA run. The visual representation of data transformation and preparation for final reports are shown on Figure 5.

5. BI Tools for Statistical Data Analysis

A tool for data analysis and reporting used in SSO is SAS Adjustment analysis tool, Corrections and Reporting tools which are the parts of SAS EIS software tool. It is powerful and user-friendly software with many possibilities provides many-phrasal data and much table analysis (Figure 6).

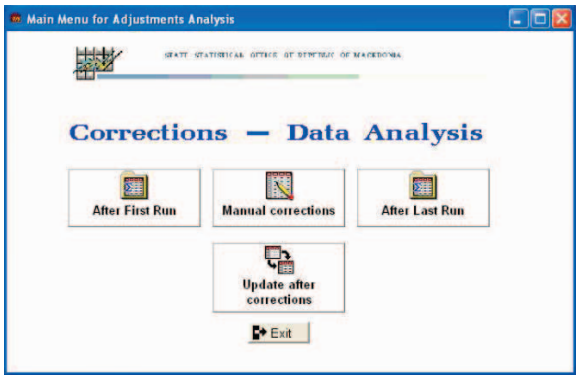

Figure 6 SAS Corrections – Data Analysis screenshot

After previous phases of data preparation, the analyst of each institutional sector in the National Accounts sectors makes analysis of the classification tables (Figure 7) with Adjustment analysis - Corrections tools. There are possibilities for analysis after the First run phase, making some manual corrections, after last run analysis and some other updates and corrections. These analyses are made with the selection of desired data from dimensional tables of DW for couple of classification codes. The experienced analysts from the NAS can explore data in each phase and can find some irregularities, if there is something like this. After this analysis, data must be saved in the final fact and dimensional tables and can be used for the next step – data dissemination. The final data can have some correction, but they must be strongly controlled from the NAS analysts.

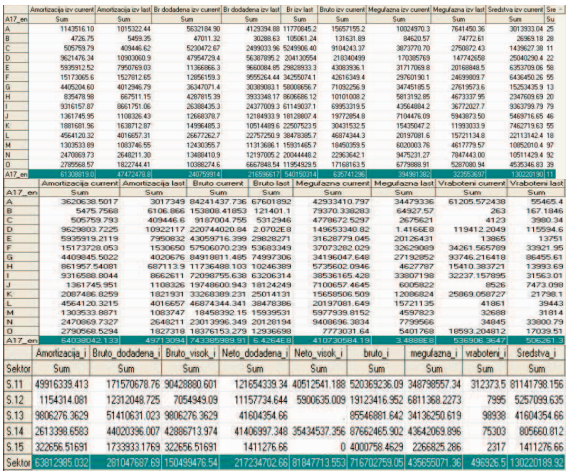

Figure 7 – Tables compared after firs and final run phases

Data dissemination is made with another Adjustment analysis – Reports tool (Figure 8 and 9). This tool provides many types of reports, depending of analyst's need and desired data. The analyst can provide analysis of integrated accounts, selected accounts or other Additional analysis depending of its demands. Reports can be enlarged if there are some demands from NAS.

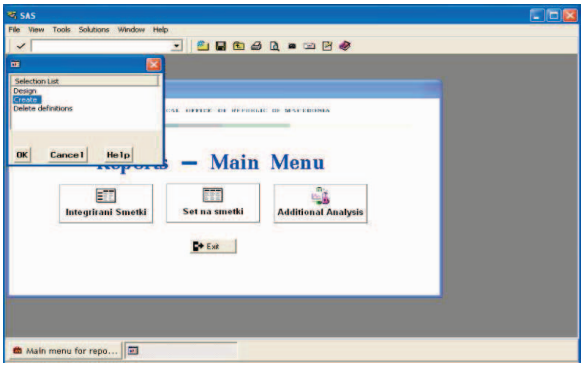

Figure 8 – Adjustment analysis - Reports tool

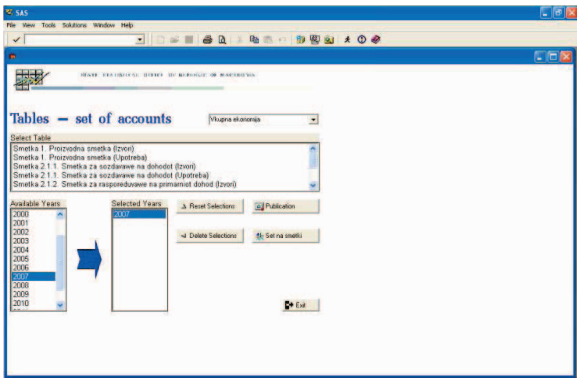

Figure 9 – Selection of parameters with Reports tool

This module can be enriched with SAS Graph tool, but it isn't implemented yet. In this moment, the data dissemination is made only with the table's reports. But, this is an excellent base for visual data analysis with other software tools

6. Tools for Visual Data Dissemination

However, the information dissemination isn't DW administrator task, but it is specific task which is dedicated to the SSP sector for data publishing. Some of the annual editions cover printed data dissemination. Although printed editions, SSO possess web site with embedded databases and possibility to browse desired data from this web site. The web site is made in the Microsoft ASP.NET platform and has a possibility to combine and present selected data from the DW. The user has a spectre of tables from which one can pick desired information and parameters for statistical analysis. The human-computer interface is easy to use and permits data visualization in the form of business graphics or maps. Some examples of use of these HCI are shown in Figures 10 (a, b, c and d). This example presents some statistic selections of data from the population in the Republic of Macedonia.

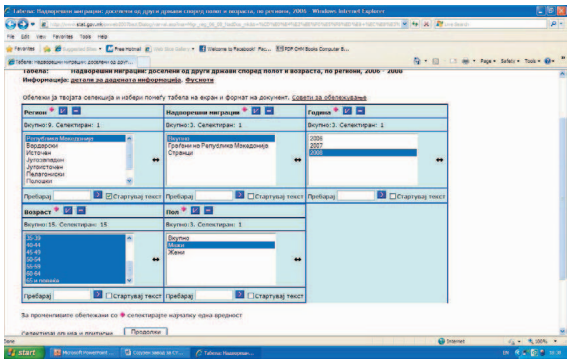

Figure 10 a – Web selection screen from SSA database

The web site is compatible with EU standards for statistical data representation. Data dissemination can be reached via web browser or PC-Axes for Windows. This easy HCI interface provides a wide range of data combination from the SSO Data Warehouse. The data can be viewed as tables or visualizations like business graphs or mapped data in the Macedonian maps. Also, they can be saved as Excel files in the user's desktops, prepared with other software tools or import in the company's databases. The first selection is

population, foreign trade, population State inventory, agriculture and regional statistics. Depending on the selected thematic area, the application selects data from multidimensional tables of DW, the selected category and data is collected in temporary tables and can be presented in the selected visual format, graphics or maps with graphical data presentation (Figure 11 a and b).

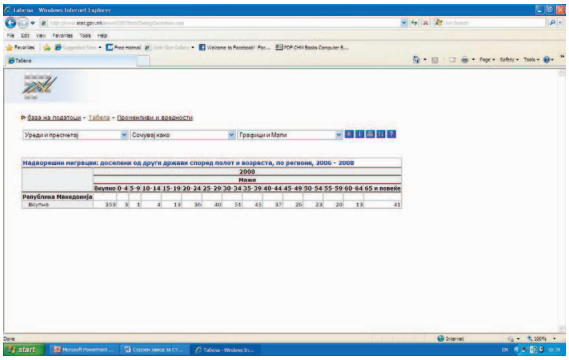

Figure 10 b – Selected data, presented in table format

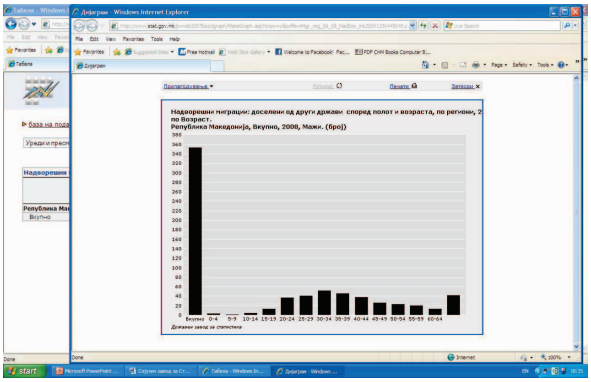

Figure 10 c – Data from 10 b presented in a visual format

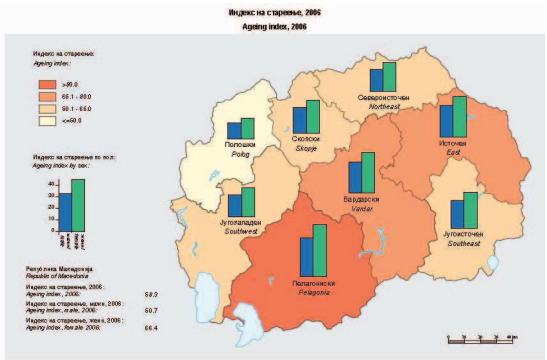

Figure 10 d – Combination of Map and BI graphics (coloured bars)

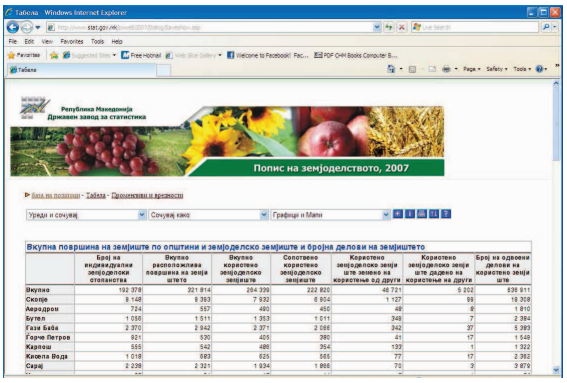

Figure 11 a - Agriculture Data analysis for selected data in a web table format

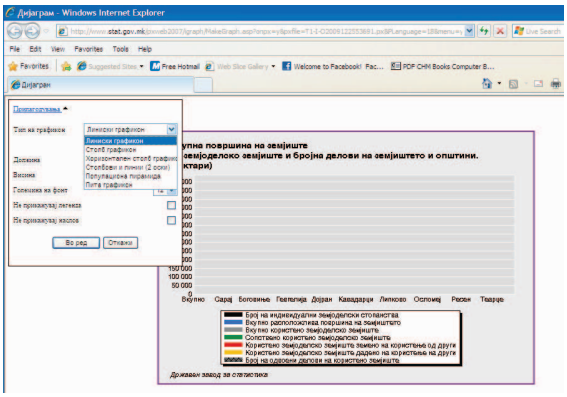

Figure 11 b Agriculture Data analysis and selection of desired visual format for data representation on the web

7. Conclusion

SAS DW is a base which is used for gaining very complex statistical calculations for the need of SSO and official statistical data, produced in the Republic of Macedonia. They use data from different sources, under different platforms and systems. The data must be very carefully cleansed, transformed and loaded in the target destinations in multidimensional tables in DW. After that, the data is the base of calculation of the sector's data in aggregated tables, depending by the statistical logic, nestled in the Excel file. These files are the fundamental databases for data dissemination for public usage via SSO web file and user friendly web software tools for data selection and data visualization. It is a very successful story for gaining BI information via statistical calculations from integrated state data. The data is disseminated via SSO web. It is a very user-friendly web site with a wide range of possibilities for statistical data representation.

7. References

[1] Kozielski S., Wrembel R., New Trends in Data Warehousing and Data Analysis, Springer, 2009, [Pages 93-113];

[2] Ponniah P, Data Warehousing Fundamentals, A Comprehensive guide for IT Professionals, A Wiley - Interscience Publication, 2001, [Pages 111-130];

[3] SAS Institute Inc., SAS Data Integration server, USA, 2009; www.sas.com/offices, [2009/2010];

[4] Cartier J., Painless Graphics: The Click, Drag and Drop Approach of Graph-N-Go, SAS Institute Inc, 2009

[5] Karp A.N\H., Working with SAS Date and Time Functions, Sierra Information Services Inc, 2008;

[6] http://www.sas.com/rnd/webgraphs/, [2008/2009]

[7] http://www.stat.gov.mk/[2009/2010]