Transforming administrative data to statistical data using ETL tools

Paulina Kobus, Paweł Murawski
Central Statistical Office, Poland,
P.Kobus@stat.gov.pl, P.Murawski@stat.gov.pl

Abstract: This paper focused on Administrative sources and measures of data, ETL tools as instruments of transformation of administrative records in statistical registers. It presents develop its various stages, starting from the extract data. Indicate the sample registers processed by the public statistics. The most widely developed part of the paper will be issues related to data transformation and the transformation of public registers into the statistical register. Then briefly discuss the final stage of the ETL process – which is loading. At the end, the summary will focus on the problems and difficulties associated with transformation of data and the benefits of administrative data for statistics.

Keywords: data integration, ETL tools, statistical register

1. Introduction

The aim of this paper is to provide processing of data from administrative sources, as part of the ETL process, covering all activities in the data sets in such a way as to obtain a result of statistical register, which is a complete set of data allows to carry out research in official statistics.

2. Extract data

In the section on extract data is presented load process to the database and work on consolidation data from various source systems, extract data into the production environment based on the SAS software and converting data into one format that is suitable for processing – SAS tables Extract data into the production environment based on the SAS software.

The first stage of work on datasets is to extract them and put into the production environment based on software from SAS Institute. We use the application called Data Integration Studio, as well as Enterprise Guide.

Obtained data has various format, for example txt,. xls, .csv, xml, MS SQL databases. Import means to consolidate data from various source systems and converting data into format that is suitable for processing - SAS tables.

An integral part of the import is to check the correctness of the data and its structures. These include in particular, the number of imported records (if agrees with the number of records submitted by the provider of information) and verify the correctness of assignment of data to individual columns (that means checking if text values contains the text, if the length of the field is suitable for data variables, etc.)

3. Transform data

Data transformation means a series of activities in the production environment consisting of: profiling - the creation of a report on data quality, unification/ standardize data, parsing (separation) or combining variables, standardize with schemes, conversion, validation, deduplication, data integration.

When dataset is successfully extracted, a process called profiling take place. We create a profile / report of the quality of data, so we can check (at the level of numerical and percentage) the rate of errors for each variable in the set. In profiling, we can obtain information about the number of completed records, the number of unique entries, patterns and incorrect data.

The next step is data standardization – unification and brought down to a defined standard – values occurring in certain columns.

Parsing is a separation of variables - for example, the division of one column 'address' on columns: 'street', 'town', 'home number' or partial name and surname from one text field.

Table 1. Example of standardization data

Incorrect data format	Format after standardization
1985-02-21	19850221
1985.02.21	19850221
1985 02 21	19850221

The example shows variable date - before and after the unification.

Voivodeship	City	Street	Place of birth
MAZOWQIECKIE	WARZSWA	ul. DŁUGA	LONDYN - ANGLIA
MAZPWOECKIE	WARS-AWA	Ulica DŁUUGA	LONDYN – WLK BRYTANIA
ZAZOWIEVCKIE	AWRSZAWA	DLUGAA	LONDYN/CHELSEA
MZAOWIECIE	WARSZAAAWA	DŁUGA (ul.)	LONDYN BRIDGE

Table 2. Example of standardization data

Voivodeship	City	Prefix	Street	Place of birth
MAZOWIECKIE	WARSZAWA	UL	DŁUGA	LONDYN

The above example illustrates effect of parsing and also standardization variable ‘street’. After this processes, incorrect values are replaced by the correct values.

The next step transform data is validation. The validation is a process of checking the correctness of data and correcting abnormal values according to the algorithms prepared by methodologists. Sometimes it is also necessary to exclude from further processing records, which improvement is impossible. Through this process we are able to obtain better quality of data.

Validation is performed on the datasets already pre 'cleaned' in the previous stages of work.

Another action is data deduplication. Deduplication is the process of removing repeating units and merge the information in the same records. It requires a detailed analysis, often including legal acts analysis. It is individual for each register. As a result of deduplication – we obtain one unique record of all the possible and unique information.

One of the last action is aata integration is a process of selection the best, most current and correct value of several or a dozen of registers. It is a process to create a statistical record, which will be available for use by analysts.

3. Loading data

Statistical register is transferes from the production area to the analytical environment. In this process it is important to use mechanisms for quick loading large amounts of data. In the analytical area further work on the data goes on, such as production of summary tables or generate reports.