Distr.
GENERAL

WP.20
6 May 2011

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2011)
(Luxembourg, 23-25 May 2011)

Topic (iii): Innovation and related issues

Statistical data editing near the source using cloud computing concepts

Invited paper

by Georges Pongas and Christine Wirtz, Eurostat

I. Introduction

1. Statistical data editing can be seen as an iterative mechanism to improve data quality by finding and, where necessary, correcting erroneous or highly suspect data. This often requires access to auxiliary data such as historical data, data from other statistical domains or even data produced by different statistical agencies. Considering that data correction and data quality improvements could require contacts with survey respondents, data editing should be handled near the source, i.e. by the primary statistical agency, which is closer to data providers than 'secondary agencies' such as Eurostat, the OECD or federal agencies. Application of common standards by the primary agencies plays a central role in ensuring that data are treated consistently by 'secondary agencies'. In this case, such standards would cover editing rules and auxiliary data.

2. Cloud computing is a development of a variety of technologies that enable organisations to alter their approach to building IT infrastructure. There is nothing fundamentally new in any of the technologies that make up cloud computing. A cloud can consist of both software and infrastructure; it can be an application accessible via the web or a (data) server which users access when they need it. The basic characteristic of a cloud is its accessibility via a web browser or web services.

3. This paper presents the problems and proposes a (cloud-oriented) solution for statistical data editing to provide EU Member States with common rules, common metadata and auxiliary data from the past or from other EU Member States. The basic features of the solution are the use of web infrastructure, including web services, the capacity to adapt to national needs and the distribution of data and programs to multiple environments.

II. Statistical data editing near the source

4. Statistical data editing activities are of paramount importance for statistical agencies. Subject-matter knowledge and complex software are combined to perform this resource-consuming task. Data editing can be classified as interactive or batch.

(a) Interactive editing near the source is associated with internet data collection systems and, in general, computer-assisted data collection systems. Data records are accepted only if they are considered correct, unless missing values are replaced by default values for sensitive or unknown answers. This kind of interactive editing falls outside the scope of this paper.

(b) Batch editing near the source is of interest in the following cases:

- In large, high-frequency business surveys, such as Intrastat, statistical agencies apply predefined checks. In the event of errors, they either impute data or contact the respondents for corrections.
- During exchanges of data between statistical agencies, the receiving organisation applies a set of checks. In principle, these are not followed by an imputation phase but by a request to the sender for improved data.
- When a statistical agency needs to apply a complex editing procedure but does not have the appropriate software, which is available to another agency.

5. In all the cases described above delivery of correct data is slowed down by communication or transmission delays. These can be greatly reduced if the respondents or sending agencies themselves trigger the checking procedures using common shared rules and/or tools.

III. Brief introduction to cloud computing and service-oriented architecture

6. Cloud computing enables statistical agencies to increase their computational capacity without having to buy new hardware or software licences or to hire or train new personnel.

7. Users of cloud services do not have to own the infrastructure or the software. Instead, they can gain seamless access to multiple servers without knowing which or where they are located. Depending on the number of people using the service at the same time, the cloud managing software will channel the demand to the appropriate server.

8. Cloud infrastructure is classified either by location or by the services offered.

(a) Based on the *location*, clouds are classified as:

- **Public**. A public cloud is hosted by a commercial vendor. Users have no visibility or control over it.
- **Private**. A private cloud has dedicated hardware and software and is used by a single organisation. Obviously, private clouds are more secure than public ones.
- **Mixed**. A mixed cloud offers the possibility to use a private cloud for the organisation’s critical or sensitive applications and a public cloud for the rest (in the case of statistical agencies, the dissemination systems would be located in a public cloud).
- **Community**. A community cloud shares the infrastructure between the members.

(b) Based on the *services offered*, clouds are classified as:

- **Infrastructure as a service (IaaS)** which offers storage and database hosting (e.g. Amazon);
- **Platform as a service (PaaS)** which offers a development platform (e.g. Google);
- **Software as a service (SaaS)** which offers a complete application ready to use (e.g. Google gmail and Google docs).

9. Cloud applications are based on a service-oriented architecture (SOA). SOA is a way of linking resources, i.e. applications and data, on demand to achieve the desired results. The results can be delivered to final users or other services.

10. In an SOA environment, resources are made available as independent services that can be accessed without knowledge of their technical implementation. Service-oriented architecture is not tied to any specific technology but applies the following principles:

(a) It minimises dependence between services;
(b) Communication between services is defined clearly and properly documented;
(c) Services hide the logic implemented from the external world;
(d) Application logic is divided into services with the intention of promoting reuse;
(e) Services can be put together to form composite services;
(f) Services have control over the logic they encapsulate;
(g) Services retain minimum information specific to an activity, i.e. they are stateless;
(h) Services are designed so that they can be found and accessed via available discovery mechanisms.

11. Given the nature of the activities of statistical agencies, it could be interesting to consider combining private or community clouds with SaaS. These cloud variants permit reuse of existing statistical software. Special attention has to be paid to data confidentiality and privacy. For good reasons, organisations are uncomfortable with the idea of storing data and applications on shared infrastructure they do not control. The need for trust in any cloud offering statistical services makes it crucial to cover, in an optimum way, issues such as data access, authentication and identity management.

IV. Statistical data editing in Eurostat

A. Introduction

12. Eurostat is mainly a secondary data collector receiving data from national agencies. Hence, it is rarely near the respondent. Only in very few cases are primary data collected by Eurostat (air transport statistics are one example of such an exception). In practice, production of statistics in the European statistical system (ESS) is shared between EU Member States and Eurostat. In general, primary data are collected and processed in the Member States. Microdata or tabular data are subsequently transmitted to Eurostat. Eurostat checks and validates the data received from ESS members before further dissemination and processing, such as calculation of European aggregates.

13. The volume of information consists of more than 61 000 files per year. The formats are Excel, Gesmes/TS, SDMX/ML, but also various delimited or fixed length types. Assuming that reading and editing takes, on average, two hours per file, the resources add up to more than 70 person/years.

14. Given also that, institutionally, Eurostat rarely imputes erroneous data (and is usually not in direct contact with the respondents), but sends error reports back to the national authority instead, the current editing set-up is not optimum.

15. Awareness of this situation prompted Eurostat to seek a solution which keeps the advantages of [simulated] editing activity near the source but combines them with those of a centralised approach [in terms of data, metadata and programs] without increasing the total burden on the national authorities or restraining their autonomy in any way.

16. The solution described below makes available to the national authorities a set of web services permitting them to write their own programs which can be executed on Eurostat premises, using Eurostat’s or

national databases and metadatabases and, ultimately, a web application and interface allowing them to run Eurostat's or national editing scripts on their own data or on mixtures of Eurostat and national data.

B. EBB (edit building block), a new web-enabled editing system

17. EBB is a generic data editing, deterministic imputation, file manipulation and computation system. EBB comes in two versions with the same user interface:

(a) a stand-alone version running autonomously under Windows; and
(b) a client/server web-based version.

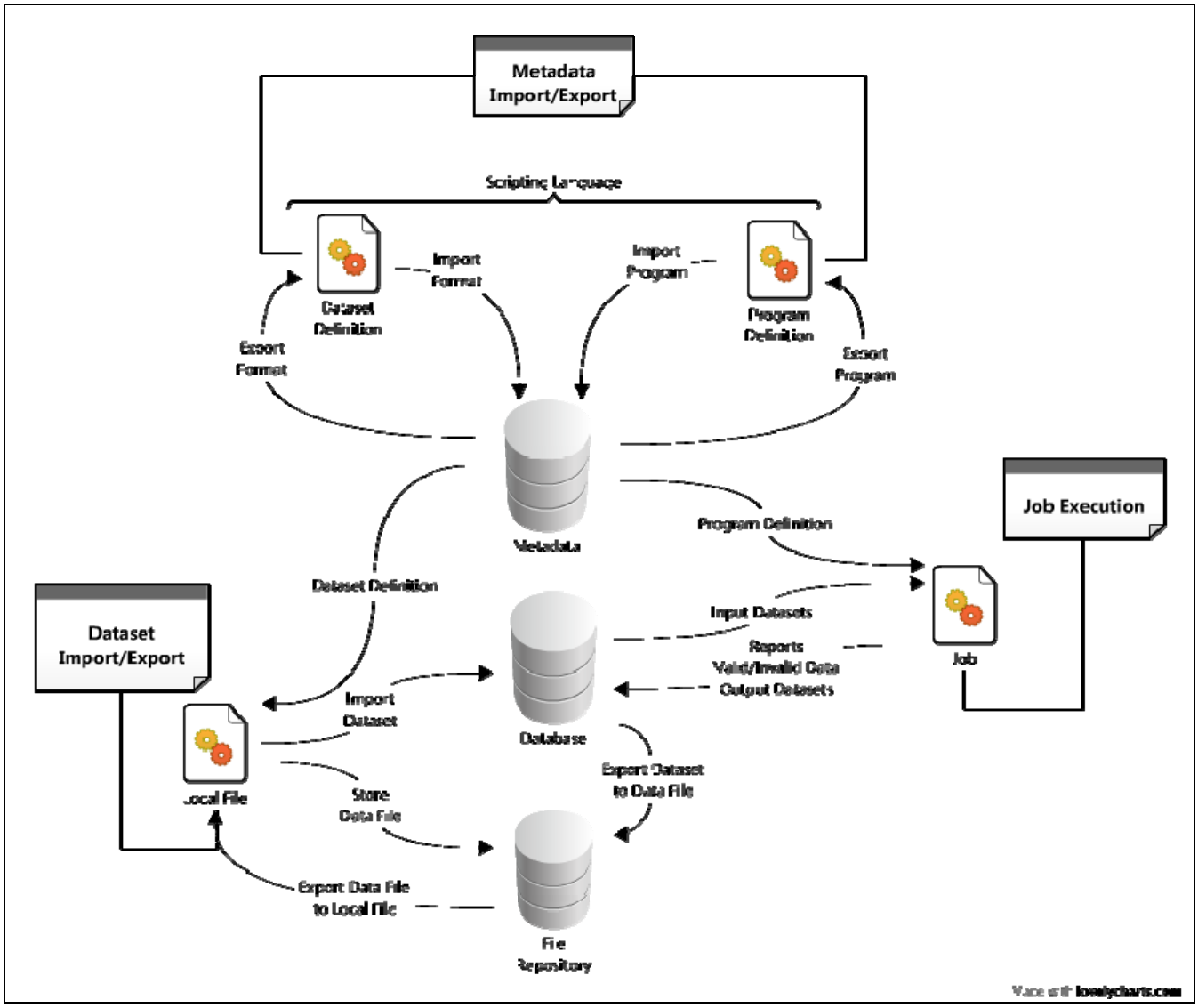
Other features include the possibility to call up the core functions of EBB via a command line interface or web services. The latter allow existing applications to call up EBB to perform validation functions.

18. EBB uses the following software components:

(a) ANTLR (Another Tool for Language Recognition) for EBB parser generation;
(b) Java;
(c) Tomcat (the stand-alone version) or weblogic (the client/server web-based version);
(d) Hibernate for the database operations;
(e) Postgres database (for the stand-alone version) or Oracle database (for the client/server web-based version).

19. EBB is an entirely parametric application which makes no specific hypothesis on the shape, size or number of files or location of data.

20. The diagram set out below illustrates the architecture used:



C. Functions of EBB

21. EBB offers the following functions:

(a) Possibility to run in stand-alone or client/server mode;
(b) Separation of the developer's and end-user's interfaces;
(c) Support for categorical, text or quantitative variables, the number of which depends on DBMS limits;
(d) Support for arithmetic or logical rules, application of which can be preceded by a condition (for example, if a person is jobless then his or her salary must be zero);
(e) Support for vertical-type rules (multi-record rules), such as account consolidation checks;
(f) Multifile rules using historical data or look-up tables;
(g) Multiple data level rules for use in the hierarchical survey data;
(h) Dataset operations cover the various types of relational joins, dataset unions, dataset interleaving and aggregation by grouping;
(i) Creation of new variables conditionally or unconditionally;
(j) Various sophisticated functions such as constancy of a variable, unity of keys, various functions for missing observations, substring operations, etc.;
(k) Outlier-oriented computation implementing the Berthelot-Hidiroglu method, Sigma Gap method and time series outliers, following the Tramo Seats methodology;

(l) Import/export of data, metadata and program scripts.

22. At the end of execution of a job, EBB produces a cluster of information which includes:

(a) The data subset containing records which failed at least one rule;
(b) The correct data subset;
(c) A statistical report on the errors;
(d) A link dataset linking failed rules and original records. Each instance of a failed rule may be followed by the values of the variables implicated;
(e) The output datasets generated by the various dataset operations such as aggregations.

D. The EBB metadatabase

23. The EBB metadatabase is the core of EBB. It consists of about 30 relational tables concentrating information about:

(a) The users in terms of data accessible, roles (developer and/or final user) and activity history;
(b) The various data formats used in all the statistical domains;
(c) Location details of the external data to use (file names and other relational databases);
(d) Parametric SQL statements for run-time data access;
(e) Parameters and their values for program calls;
(f) Program script information. A program script is a set of named steps, each of which contains individual commands. A script does not make direct reference to any dataset but only to variable names. Dataset information is inserted at run time, using the parameter information;
(g) The results of program execution, such as job status, names and location of the results;
(h) Indexing and partitioning information, necessary for vertical and hierarchical information.

E. EBB web services API (application programming interface)

24. In order to offer the possibility of using EBB not only as an autonomous application for statistical editing but also as part of other applications, its development followed a layered approach. First, the function was broken down into about 70 activities (the API) and then this API was used to develop the EBB interface. Currently in most cases the EBB is used via its API and not its native interface.

25. The API function is organised as follows:

(a) **Administration services**, dealing with data domain discovery and data management, user and role management and user access management;
(b) **Program services**, dealing with script, parameter and data import/export templates and data format management;
(c) **Job services,** dealing with execution initiation, execution follow-up, job deletion and job results storage and retrieval.

V. Conclusion

26. Eurostat has developed an application for editing and deterministic imputation which not only uses web technologies but also makes available the underlying programming tools permitting other applications and users in Eurostat or national statistical agencies to use it. An application of EBB is a set of accessible metadata and a single-script execution engine and storage facility allowing calls from anywhere in the network and using data located anywhere in the network. EBB was conceived to maximise ergonomy, scalability data and metadata unification. The hope is that use of EBB will increase data quality and result in streamlining and efficiency gains.