UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Working paper
27 February 2015

Workshop on the Modernisation of Statistical Production
Meeting, 15-17 April 2015

Topic (i): Business and IT changes that will impact statistical production

FROM DATA STORE TO DATA SERVICES - DEVELOPING SCALABLE DATA ARCHITECTURE AT SURS

Prepared by Tomaž Špeh

Tomaz.Speh@gov.si

Statistical Office of the Republic of Slovenija

Summary

A statistical organization faces tough demands for making its production more efficient while at least maintaining or preferably increasing quality. These demands are directed towards all phases of statistical production. In the most recent years, the Statistical Office of the Republic of Slovenia (SURS) has performed substantial part of the development to modernize the data collection and data processing process and its underlying IT systems.

Consequently, SURS operates a large data store, with the widest and deepest data access; furthermore, big data are becoming increasingly important as additional data sources. On the other hand, the advancement of the SOA architectural pattern, assisted by evolving common principles and frameworks designed to promote greater interoperability (CSPA, GSIM, GSBPM), encourages modernization by developing reusable and sharable software components for the statistical production process.

Relational data warehouses have long dominated analytics, reporting, and operations. However, their slow-changing data models and rigid field-to-field integration mappings are too brittle to fulfil future needs.

Data virtualization separates the physical source of data from its potential applications, with the goal of increasing business agility. Instead of physically grouping and storing datasets together - as in one big giant database or data warehouse stored on a big server or cluster - virtualized data are stored in a distributed manner, across a series of disparate servers. A logical, unified, and federated view of the data is then created, using the data's metadata, providing streamlined access (via Web services or straight SQL) to data objects.

This paper summarizes SURS's experiences and describes different use cases where data virtualization simplifies further modernization of statistical production.

I. Background

1. Statistical authorities are under pressure to produce data faster and at lower cost, to become more responsive to users´ demands, while at the same time providing high quality output. One way to fulfil this is to make more use of already available data sources. The Statistical Office of the Republic of Slovenia (SURS) has been moving towards an increased use of administrative data sources for statistical purposes as a substitution and/or as a complement to information previously collected by sample surveys. Administrative data sources are similar in structure, but are not the result of a sample survey. They are typically collected in support of some administrative process. The variables are not chosen or defined by the NSI, as opposed to the variables in sample surveys. Alternative data sources are becoming increasingly important. They are not comprised of a set of records directly corresponding to units in a target population. These kinds of data sources often register events. Such data can be generated as a by-product of some process unrelated to statistics or administration. Since these data files are often much larger in size than sample data or administrative registers, the term 'big data' is sometimes used in these cases.

2. Within SURS there is a growing need to modernize statistical data collection and processing. Key words are decreasing costs and administrative burden, increasing efficiency and flexibility. We are investigating ways to disclose all kinds of new data sources that become available through the global use of modern technologies such as the internet, mobile phones, automated scanning techniques, etc. At the same time, new technologies offer broad new possibilities to modernize statistical data processing processes.

3. SURS collects data in different ways and in different formats. To ease the collection of business surveys, web collection system ESTAT has been developed. For some administrative data sources with frequent data exchange, the direct database replication method is used; the majority of others come into the office through the Government Secure Data Exchange Hub or SFTP server. Recently, the need for supporting efficient acquisition of alternative data sources (for example scraped data, mobile data, scanned data) has emerged. The new technologies bring with them new challenges in such areas as efficient processing, integration into a multi-source environment, privacy and security issues, and cooperation with partners outside official statistics.

4. Statistical data processing has always been a demanding, time consuming and consequently quite expensive task. To overcome or at least reduce the gap between the above mentioned demands, in recent years a lot of effort has been put into the rationalization of statistical data processing. As a consequence, SURS is developing a modernized statistical data processing system consisting of a few modules which aim at "covering" the different parts of the statistical process (e.g. data validation, data correction and imputation, aggregation and standard error estimation, tabulation).

5. One of the challenges in this process of change is the integration of sources and collection modes and following to that, the standardization of collection methods and technologies. Besides this, a second and probably even bigger challenge is the integration of the collected data into the statistical production: How to make optimal use of all available data sources (existing and new)?

II. Scalable statistical data architecture

A. Statistical data warehouse at SURS

6. Building a statistical data warehouse is considered to be a crucial instrument in this process of change in the integration of sources and collection modes and following to that, the standardization of collection methods and technologies. This approach can be used as a tool that helps to identify the particular phases and elements in their process of statistical production that must be common/ reusable. There are several ways of defining the statistical data warehouse: a strong focus on data access and output or also process integration (process driver), static, data storage or dynamic, data flow? But in all various concepts the goal is the same: to create a central data store, integrating new data sources and statistical output. When developing statistical data warehouse, design principles to provide the right enterprise environment should be followed.

7. Data from everywhere need to be accessible and integrated in a timely and secure fashion. Unlike internal systems that IT can use to manage data quality, many of new data sources are incomplete and inconsistent forcing statisticians to leverage the data warehouse to clean the data or synthesize them for analysis. Advanced analytics has been inhibited by the difficulty in accessing data and by the length of time it takes for traditional IT approaches to physically integrate them. The data warehouse needs to enable statisticians to get the data they need in a timely fashion, either physically integrating them or accessing virtually-integrated data.

8. Building solutions must be fast, iterative and repeatable. Statisticians need to get data and create tentative models fast, change variables and data to refine the models, and do it all over again. The data warehouse needs to be architected to ensure that solutions can be built to be fast, iterative and repeatable.

9. The statisticians need to “run the show”. IT has traditionally managed the data and application environments. In this custodial role, IT has controlled access and has gone through a rigorous process to ensure that data are managed and integrated as an asset. The NSI, and its IT, need to entrust statisticians with the responsibility to understand and appropriately use data of varying quality in creating their analytical solutions. Data are often imperfect, but statisticians are the trusted advisors who have the knowledge required to be the decision-makers.

10. Sufficient infrastructure must be available for conducting statistical analyses. This infrastructure must be scalable and expandable as the data volumes, integration needs and analytical complexities naturally increase, taking into account governance and privacy policies. Insufficient infrastructure has historically limited the depth, breadth and timeliness of advanced analytics as statisticians often used makeshift environments.

11. SURS will continue to drive the demand for data warehouse centric solutions, as arrays of statistical applications rely on data sourced from data warehouses. However, business changes often outpace the evolution of data warehouses. And while useful for physically consolidating and transforming a large portion of data, significant volumes of statistical data continue to reside outside the confines of the data warehouse.

12. Data virtualization is a solution that sits in front of multiple data sources and allows them to be treated as a single SQL database and therefore has the potential to improve the existing data architecture at SURS. Unlike hardware virtualization, data virtualization deals with information and its semantics – any data, anywhere, any type – which can have a more direct impact on business value.

B. Data virtualization versus Extract, Transform, and Load (ETL) data integration

13. ETL is designed to process raw data as it bulk copies complete data sets from the source systems, translates and, often, cleanses the data to improve their quality, and loads the resultant data set into a target data warehouse. ETL is a critical part of a physical data consolidation strategy which replicates data in the data warehouse. Data replication might not be an acceptable solution in highly regulated scenarios where governance and privacy policies restrict or prevent data duplication in an attempt to control and manage data access. Consolidating data from multiple sources into a shared data warehouse requires that any source data are mapped to the schema of the target data warehouse. This in itself necessitates the upfront modelling of the different data sources into a common data model – a process that is usually long and taxing as it involves multiple stakeholders and data owners, each with their own perspective on the final data model. However, ETL does provide benefits to the implementing organization and provides a suitable platform for building a single (albeit replicated) source of data to the organization. Additionally, it improves productivity by supporting the reuse of data objects and transformations. It imposes a strict methodology on the management of the data assets and greatly improves metadata management, including better impact analysis of changes in the underlying data source schemas.

14. Data virtualization, on the other hand, abstracts, federates, and publishes a wide variety of data sources to consuming applications in an array of different formats. In doing so, it simplifies and accelerates the process of accessing, combining, and utilizing these disparate data sources and hides the complexity of the different data sources from the consumers. Data virtualization focuses on creating a unified common data model across diverse data sources rather than highly efficient data movement from data source to data warehouse. The data virtualization platform creates logical views of the underlying data sources. These data sources can be structured, semi-structured (such as data on the Internet, social media, delimited files, etc.), or unstructured data – the complexity of accessing these data is hidden from the consumers. The logical views can be combined and enriched to create ‘business’ views (also called ‘derived views’), which can then be exposed as data services to consumers in many different formats; SQL, Web Services (SOAP/XML and RESTful), and so on.

15. Providing virtual views on to the underlying data sources also allows data virtualization to be more agile when adding new data sources and changing the underlying logical data model. Data virtualization supports rapid development iterations with value being added to the solution in each iteration (e.g. every week or two weeks). This is very different from the typical ETL project that takes many months of upfront planning and data modelling before any data can be consolidated in a data warehouse and, after deployment, it is very hard to change.

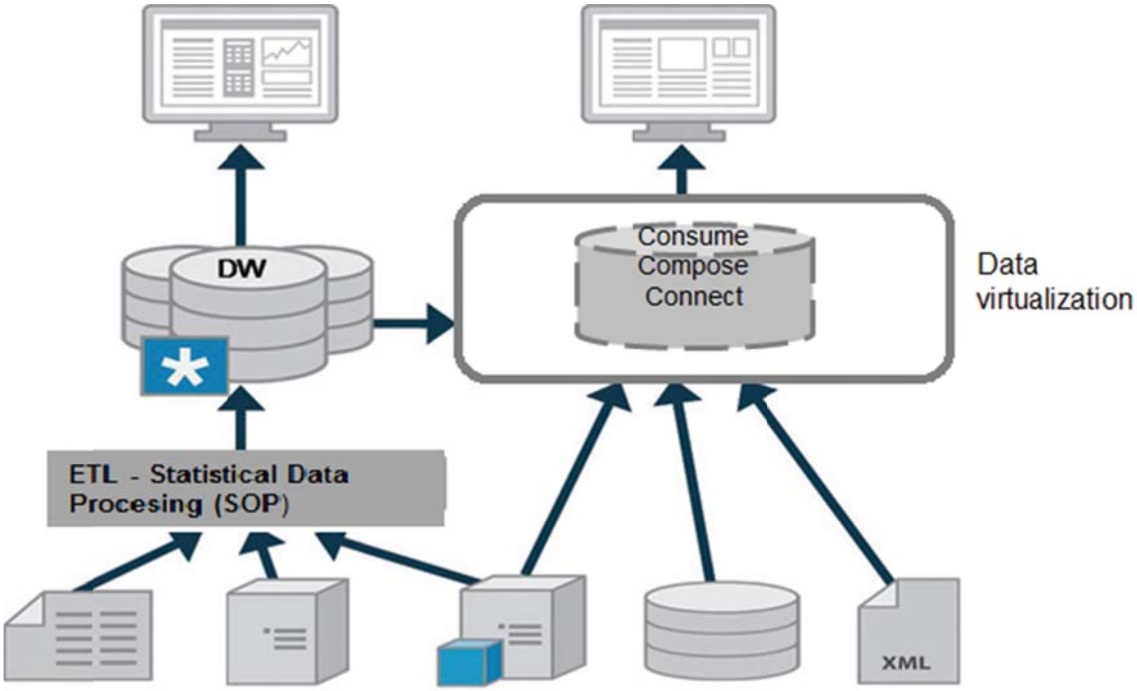
16. Today, statisticians at SURS still rely on IT-built reporting or custom extracts fed by data-integration tools, and then use spreadsheets to fill the gaps. Gathering requirements and designing and building the IT-built extracts severely slows down the time-to-solution. Data

integration, data management and building a consistent, clean and conformed data warehouse will continue to be the responsibility of the IT group. The data-integration capability will expand beyond traditional ETL to include data virtualization. The data access and integration layer needs to empower statisticians to get the data they need as quickly as possible, recognizing that getting the best available data, even if not perfect, is better than making a decision with incomplete data or by using a data shadow system.

17. Data virtualization empowers statisticians in other cases. First, it enables them to expand the data used in their analysis without requiring that it be physically integrated. Second, they do not have to get IT involved (via business requirements, data modelling, ETL and BI design) every time data need to be added. This iterative and agile approach supports data discovery more productively for both business and IT.

18. Critical data to classify, filter and analyse are often not available from office sources, but may require an external data feed. Data virtualization provides the ability to discover the data before extract and import into the data warehouse.

19. The access options – provided that security and privacy requirements are met – include query sources directly, data services, using local files and data virtualization. The first three alternatives are all point-to-point access where the statisticians must know about the source, secure access and then navigate the source. Data virtualization is an architectural option that creates a data source catalogue that can be saved, shared and documented for statisticians and augmented by the IT staff.



III. Data Virtualization – Use Cases

20. SURS is conducting the pilot project Population Statistics Project using mobile positioning data. The objective is to enhance statistics about population (active, retired, etc.), distribution regarding time and location, and population mobility using mobile positioning data. Anonymized raw (micro) data have been transferred to the statistical organization for processing. According to our Regulatory Compliance, the data are considered as confidential. This project brought new challenges in areas such as efficient processing, integration into the current environment, privacy and security issues. A so-called data lake – a storage repository that holds raw data in its native format – started to be set up using HortonWorks with JBoss Data Virtualization.

21. Considered as a Self-Service Business Intelligence, Data Virtualization provides business-friendly representation of data, allowing the statisticians to interact with their data without having to know the complexities of their database or where the data are stored and allowing multiple BI tools to acquire data from centralized data layer. It provides better insights from data sources using Data Virtualization to integrate with existing information sources.

22. In line with SURS's Regulatory Compliance, the Data Virtualization layer delivers the data firewall functionality. Data Virtualization improves data quality via centralized access control, robust security infrastructure and reduction in physical copies of data, providing security mechanism such as role based control and data masking, thus reducing risk. Furthermore, the metadata repository catalogues data locations and the relationships between the data in various data stores, enabling transparency and visibility.

23. The benefits of this architectural pattern are many, and they include both business and technical value. Data virtualization enables business agility, action ability, information speed, and information quality with rapid data integration, resulting in quicker time to solution for business information needs, more information opportunities with reach into the new types and greater volumes of data, more robust statistical analysis through more types of data and more extensive data integration, more complete information through reach to new data types and greater data volumes, better quality information that translates to business syntax and context instead of delivery in systems and data storage context.

24. The technical case for data virtualization is based on fast, efficient, and effective delivery of information by making data integration easier to achieve both in scope and timeliness of information, supporting the discovery-driven requirements and test-driven development needs of agile development projects, breaking down the barriers of integrating structured and unstructured data into a single consumer view of information.

25. SURS is currently running a project UROP which aims to improve the management of input and output data sources. The following are some possible examples for using data virtualization as part of modern statistical data integration stack that efficiently support related business processes.

26. Augmentation - Data Virtualization can be applied to augment an existing data warehouse with virtual views to meet new information needs. Unstructured data, cloud data, and real-time data integration can be implemented without extensive and disruptive changes to the core data model and ETL processes. Speed of delivery and speed of data are

accelerated with virtualization. Leveraging new and existing data sources more rapidly advances business agility. Unstructured data are integrated with structured data and new reporting and analytic applications can be more quickly implemented.

27. Federation – SURS has multiple data warehouses for a variety of reasons – mergers and acquisitions, independent departmental initiatives for data integration, etc. Whatever the causes, the result is new silos of data without full enterprise integration. The challenge is to deliver an integrated view from many different data warehouses without disrupting the independent operations of each warehouse. With virtualization, each individual warehouse operates independently, serving the users and purposes for which it is designed while simultaneously participating in federated views that support enterprise-wide perspective.

28. Disposable data marts – As demand for information accelerates, the demand for new data marts grows. Data virtualization enables the ability to create virtual data marts that can be deployed quickly, without the increased development effort and production workload of more ETL processing. Virtualization enables the concept of disposable data marts and creation of new data marts easily – a particularly powerful technique in highly volatile business and systems environments.

29. Complementing ETL – Many older data warehouses acquire, process, and load data through ETL processing, using data sources that are typically structured relational data. With technology advances, many new data sources are incompatible with existing ETL processing. Modifying existing ETL processes to access these sources is complex and risky. Data virtualization is an effective way to remove or reduce data source to ETL incompatibilities. Using a virtualization tool you can pre-process problem data sources, creating views that are readily accessible by your ETL technology. Changes to existing ETL processes, and the risks inherent in those changes are substantially reduced.

IV. Conclusion

30. Data virtualization could play a key role in modern statistical data integration stacks to cover some strategic needs. Factors such as exponential data growth, new data source types that create new information silos (NoSQL, Cloud), and the extensive use of Big Data require a new infrastructure that is not covered with traditional ETL solutions alone. The answer to the original question of “when should I use data virtualization and when should I use ETL tools?” really is “it depends on circumstances”.

31. In many cases, especially those combining structured data with unstructured data or requiring real-time access to up-to-date data, then data virtualization is a better option. In some cases, where it is really necessary to copy massive amounts of data for complex analytics or historical data marts with no concerns about data freshness, ETL and static data warehouses are still the best option. Very often, the line is not that clear. The cost of data storage, operational costs of the solution, and time to market can tip the balance to data virtualization, even for projects that have traditionally used ETL solutions. Having said this, data virtualization can often be used to increase the value of existing ETL deployments by extending and enhancing the ETL tools and processes.

32. Furthermore, when deployed in a SOA environment, Data Virtualization can provide data services to any application through the Enterprise Service Bus (ESB), which will then

work as a consumer of such services. The complex combinations and data transformations provided by Data Virtualization often exist within complex business workflows with an ESB being used to implement business logic, fetching “clean” data from the Data Virtualization layer quickly and easily, as needed.

References

Seljak, R. (2014), “Metadata driven application for data processing – from local toward global solution”, paper presented at the UNECE Work Session on Statistical Data Editing, Paris, France, 28-30 April 2014