UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
ModernStats World Workshop 2022
27-29 June 2022, Belgrade, Serbia

Integrating technological advancements into a modernised statistical production process

Abstract
Meeting new user needs on more detailed and coherent statistics as well as making use of new types of data sources requires us to modernise the statistical production process. A modernised process driven by design, data and metadata is sometimes referred to as "Statistical production 4.0". Making design changes for whole statistical areas in parallel with ongoing production means a need to be able to use a combination of parallel collection phases, process phases, analysis phases and dissemination phases at the same time. This can be seen as a cluster-based/interconnected view of the statistical production process. As a consequence of this, a more agile architecture is required to meet this change as well as to support the need for faster change. At the same time as we see this change, huge technological advancement has been made as part of the data science, machine learning, open source and cloud paradigms. How can an IT-architecture be structured to make use of the development in a way that supports the need for an updated process model for a modernised production process? Can the solutions that support the new generation of technologies be integrated into our statistical production process and provide new opportunities in how we organise the production.
This paper will describe some of the ways these advancements can be used and integrated to meet the requirements for a more agile architecture to support Statistical production 4.0.

Keywords: The statistical production process, modernisation, GSBPM, architecture, CSPA, machine learning, IT-architecture, solution architecture

Background

As part of Statistics Sweden's move toward a more process-based view of statistical production around 2007, investments were made into building shared it-services to support specific process phases. Previously, statistical production was largely supported by either it-systems built by business (mostly in Excel, SQL or SAS-scripts) or it-systems built by IT for a specific statistical program (such as .Net-based systems to support production of consumer price index).

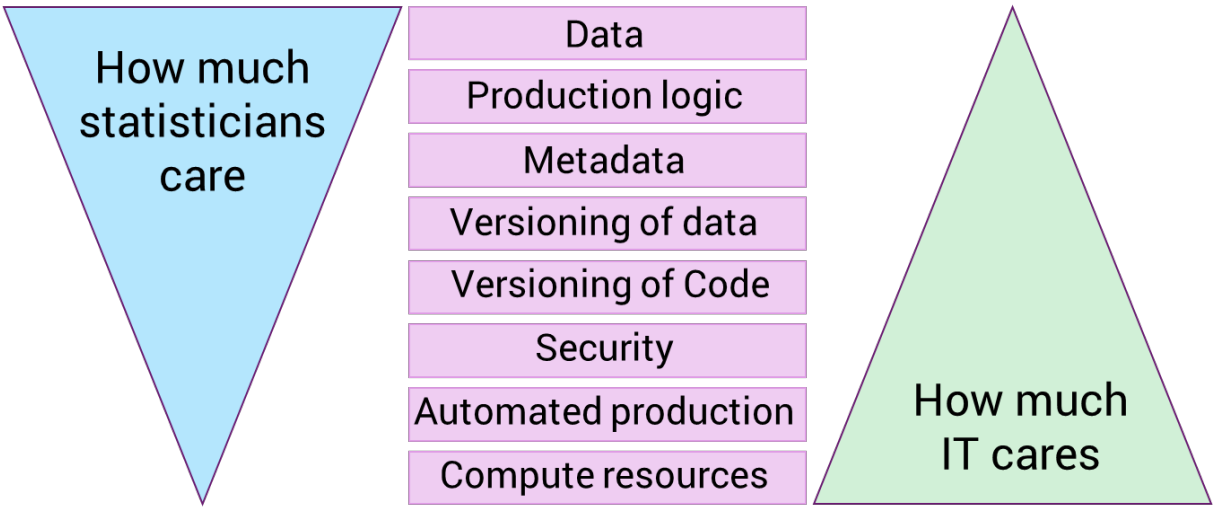
Since statistical production is supported by a large amount of production logic and since development of it-systems can be expensive, complex and take time, this is still an ongoing journey. The strategy to focus more on shared services has largely been successful to support *data collection* and *data dissemination* but has been more difficult for the *process* and *analyse* phases where the need for flexibility/bespoke solutions is high.

The move towards more shared services results in fewer accidental mistakes due to increased focus on quality aspects but come at the cost of being more expensive, complex and limits the flexibility as compared to systems built by business. Arguments could be made that this results in a decision between *flexibility, quality* and *cost effectiveness* where we are only able to reach a maximum of two of these goals regardless of strategy.

Now requirements increase to achieve an even more agile statistical production to support the need for faster changes due to new data sources being available and due to new user needs. We therefor need even more flexible it-solutions.

Flexibility vs quality

Systems built by business and systems built by IT differ in some specific ways. It is natural that statisticians focus more on translating the design/production logic to a simple it-system that can perform the necessary data processing while developers have been taught the importance of good system quality. Though not true in all cases, a simplified illustration of the focus of different professions could be:



[Inspired by a presentation by Netflix about creating Human-centric Machine Learning Infrastructure https://www.infoq.com/presentations/netflix-ml-infrastructure/]

This results in different defining characteristics of each type of it-system:

Built-by-Business:

- Design/Production logic expressed as code or as Excel sheets
- Simple database or data structures
- Development environment separate from production is either missing or a copy of the production environment
- Low focus on quality aspects such as
  - Versioning of code
  - Versioning of data
  - Security
  - Robustness
  - Documentation
  - Performance
- Flexible and quick to change
- Risk of growing too large/complex over time

Built-by-IT for specific statistical program:

- Design/Production logic expressed as parameters/configuration in a graphical user interface
- More complex/optimised database structures to handle security, versioning etc
- High focus on quality aspects
- Development, test and production environments kept separate
- Bespoke so fit for purpose but not as quick to change due to changes made by it and process of quality assurance before changes hits production
- Expensive to develop and maintain due to large amount of statistical programs/systems

Built-by-IT for many statistical programs (shared service):

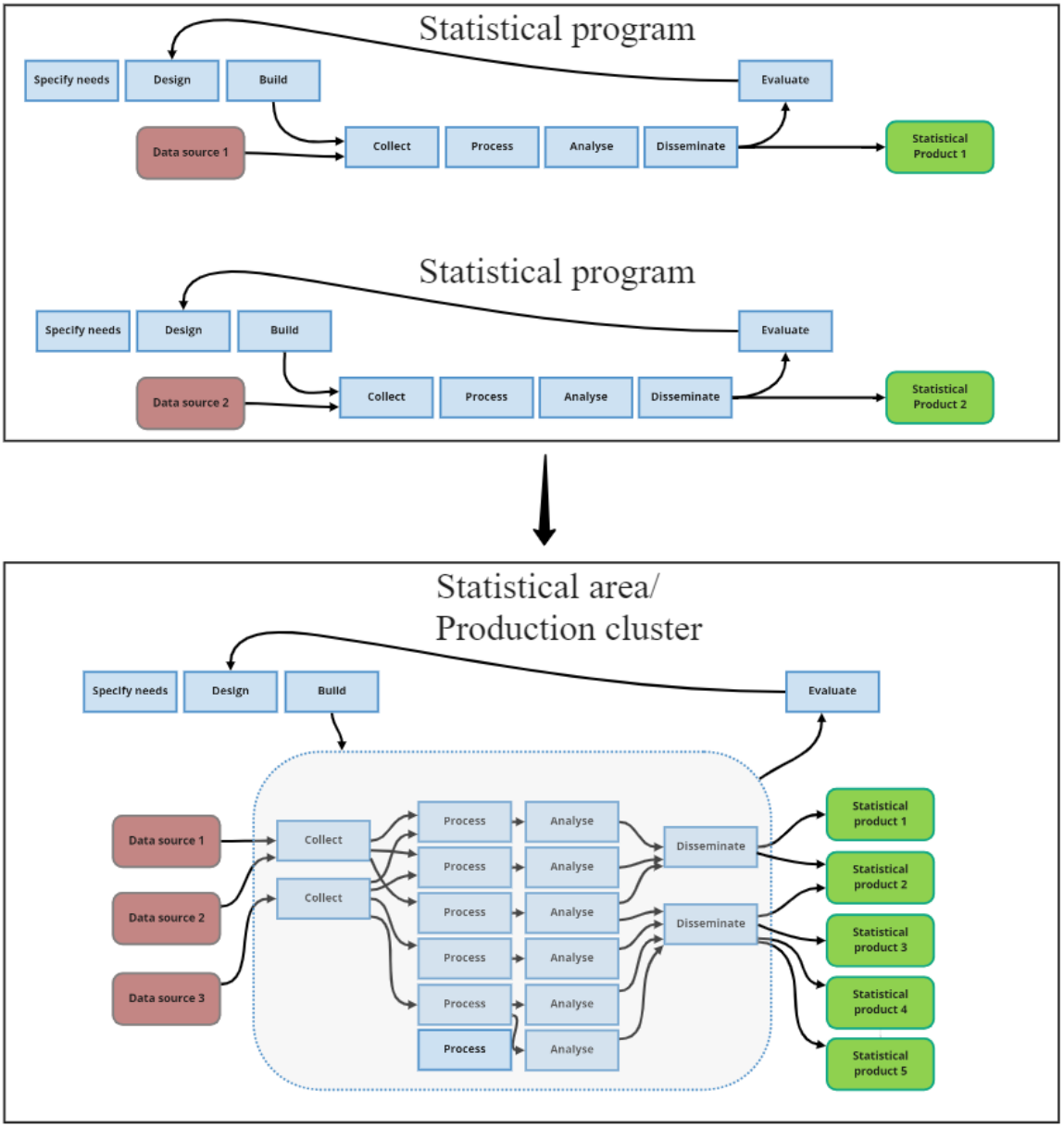
- Design/Production logic expressed as parameters/ configuration in a GUI
- More complex/optimised databases to handle security, versioning, context etc
- High focus on quality aspects
- Development, test and production environments kept separate
- Development of functionality slow due to changes made by IT and process of quality assurance before changes hits production but
- Implementation of new functionality easier
- Changes in production logic quick
- More expensive to develop and maintain than built-by-business

Though expressing design through production logic as code is not something negative in itself, the functionality for reoccurring quality requirements such as versioning, security, scaling etc needs to be delivered by the systems/platforms delivered by IT in order to achieve both flexibility and quality.

Modernisation of statistical production

The recent changes in statistical production where new data sources are integrated more frequently and new user needs are ever increasing puts even further requirements on flexibility. One way Statistics Sweden is responding to this change is to put more emphasis into redesigning whole statistical areas rather than just specific statistical programs. This means putting more focus on the design of the whole statistical area rather than focusing on each statistical program.

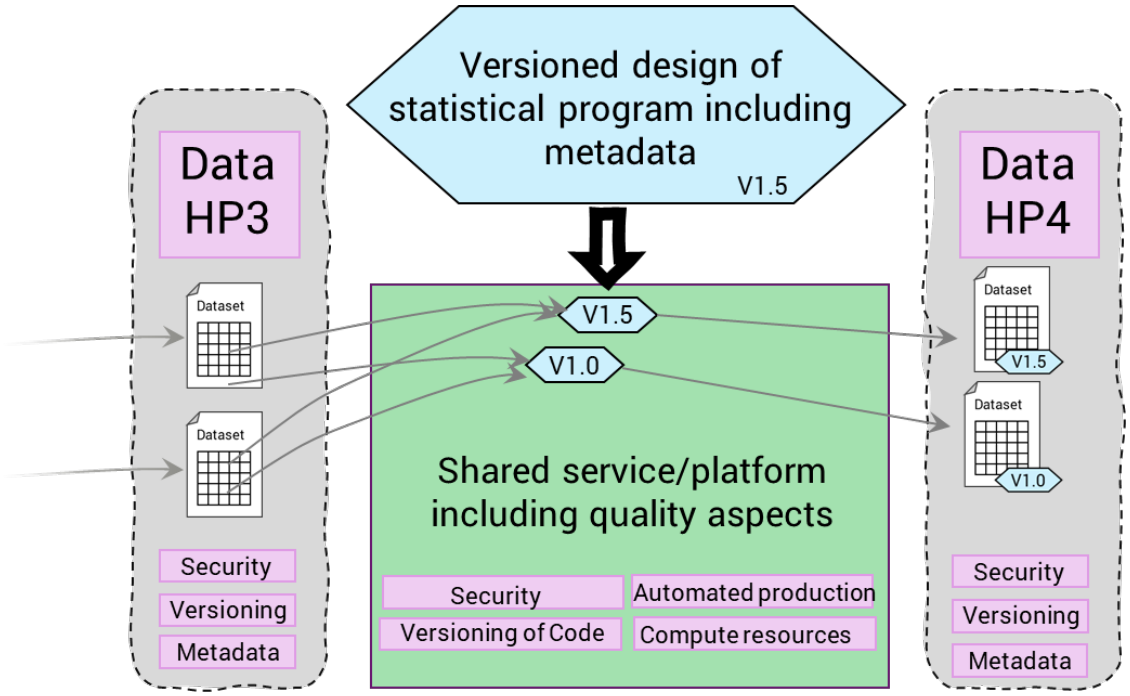
Making design changes for whole statistical areas in parallel with ongoing production means we need an architecture to support being able to use a combination of parallel collection phases, process phases, analysis phases and dissemination phases at the same time. The parallel phases could be due to different channels or data sources, but it could also be parallel as in different versions of the production logic/design to be able to try out new ways of organising the production without interfering with existing production logic. This can be seen as a cluster-based/interconnected view of the statistical production process.



To maintain control over the dataflow in a more complex, cluster-based production process, more focus is placed on data and metadata management. To achieve this, clear handover points between process steps and between statistical programs are defined. The term *handover points* is similar to the concept of *steady states* and Statistics Sweden outlines the following handover points (where security, versioning and defined metadata is crucial):

- HP 0: raw data,
- HP 1: transformed raw data,
- HP 2: treated observation register,
- HP 3: final observation register,
- HP 4: statistics
- HP 5: published statistics and data.

IT must therefore shift from delivering systems that focus on production, to deliver platforms that allow the business side to make continuous, systematic change while maintaining control over the quality aspects. This means IT develops and deploys shared services/platforms that in turn allow the business to make changes to production logic without new releases of the shared service/platform by IT. In this case the application life cycle management (deployments from IT etc) is separated from the GSBPM-phases for any *Statistical program cycle.*



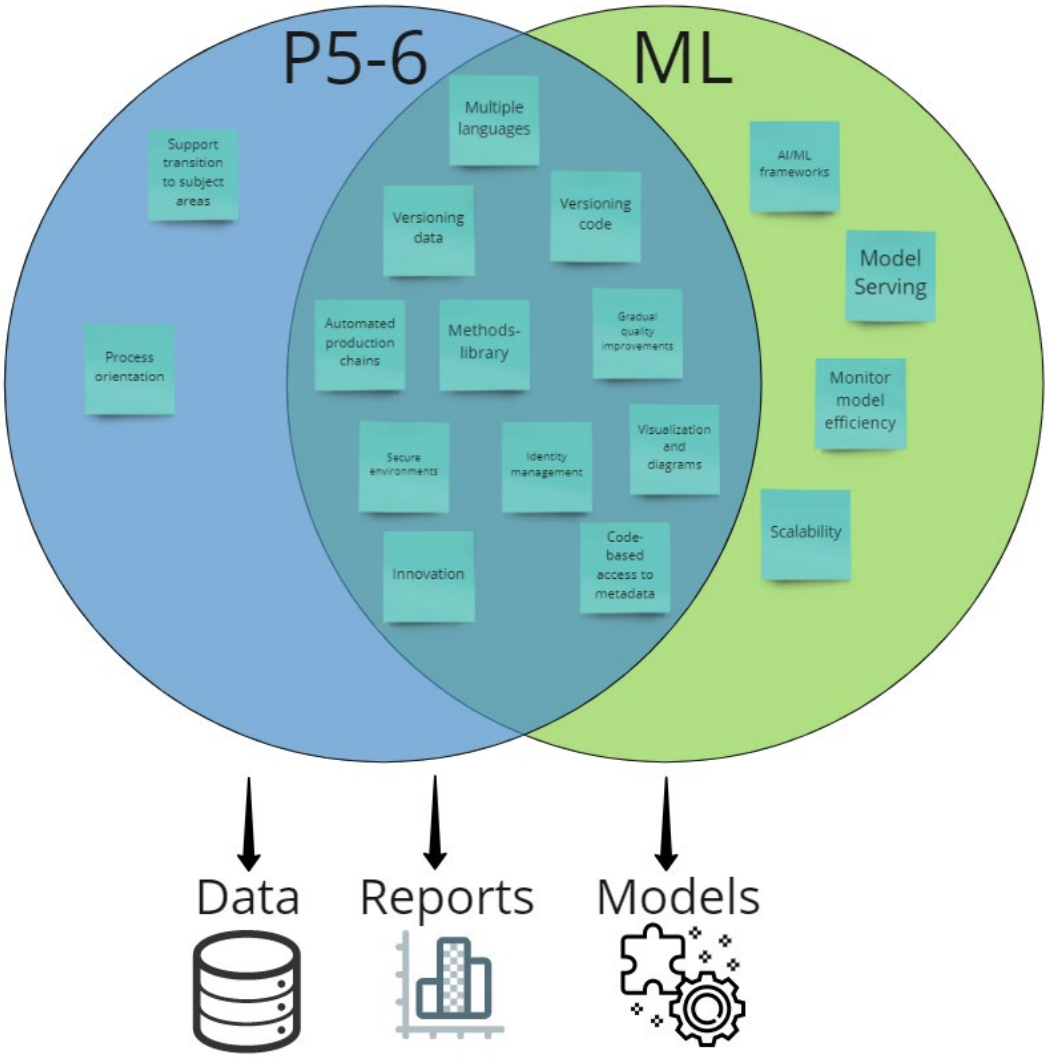
Technological advancements

As part of the data science, machine learning, open source and cloud paradigms, huge technological advancements have been made. In the area of data science, machine learning and cloud computing, the use of open source components and solutions have become very common. The large cloud providers are also some of the most active organisations in the open source community and drives the development of many of the open source platforms used in machine learning. Open source has gone from something niche to default. This has also made it possible to use a lot of cloud-based solutions on-premises.

Data science and machine learning solutions often provide functionality that are similar to the required functionality for processing data for the purpose of producing statistics. Functionality needed for both machine learning as well as processing data for statistical purposes could be:

- Versioning of data
- Versioning of code
- Secure environments
- Automated production chains
- Support for multiple programming languages
- Visualization of data
- Etc

The overlap of functionality is therefor large enough to ask the question: *Can solutions built for supporting machine learning and data science processes be a good fit for supporting traditional statistical production?*



A new initiative at Statistics Sweden (Project *BALSAM*) is aimed at investigating how these technologies can be used to deliver both the necessary functionality as well as supporting the modernisation of the statistical production process. The aim is to build a *template* for a shared service/platform which can support both traditional statistical production as well as our machine learning initiatives. Using the same type of platform for both traditional statistical production and machine learning initiatives could make these types of processes more similar and in turn speed up the adaptation of machine learning in official statistics.

The focus on using open-source solutions in this context increases portability/shareability, transparency and opens up for collaboration within the statistical community while avoiding lock-in-effects that could be the effect of using cloud-specific solutions and proprietary solutions.

The following areas are some of the ideas and topics about how we can integrate the technological advancements from the fields of data science, machine learning, open source and cloud computing into our development.

Data management

The main way of organising data at Statistics Sweden is through the use of database management systems, specifically MS SQL. Before the use of databases, flat files were often used to organise data and csv-files (and other formats) are still used quite often to transfer data between organisations since they are less dependent on specific technology platforms. For many years, databases have been a logical place to implement different functional and quality requirements such as security, versioning and efficient search. As a consequence of this, the complexity of the database models has increased in the solutions built-by-IT whereas the built-by-business-databases remain quite more simplified since accessing and writing data to a complex database model often

requires special developer skills. An example of a complex database structure is a star-based schema structure, often used for data warehouse solutions.

As a part of the technological advancements made in cloud computing, innovative solutions have been developed to manage data in new ways beyond using databases, often with built-in solutions for managing versioning, security etc. One of the commonly used technologies are so called *Cloud storage buckets* often in the form of *S3 compatible object storage* (first developed by Amazon but now available by most large-scale cloud providers as well as in on-prem solutions). These solutions manage data as objects (with the data itself, metadata and a unique identifier) without fine-grained control over data schemas etc and can therefor be used for both structured and unstructured data. Though not a solution for everything, smart implementations has made the technology popular and many *S3 compatible object storage*-solutions has smart functionality to support versioning, scalability, security etc. This means that data can, through the use of these solutions, be structured more simplified again without loosing the quality aspects that was implemented in the database solutions. This means flexible and useful data storage solutions for the business-side without loosing the quality aspects of the solutions.

Data management is one of the topics at Statistics Sweden which has proven difficult since data is, in a way, part of our DNA and it has been difficult to balance useability aspects against requirements on quality aspects. The introduction of *S3 compatible object storage* could provide new solutions to deliver on both flexibility and quality aspects.

Multi-language support, automated production chains and scaling

Statistical production often consists of a chain of functions, methods or code that together performs the different steps in the production process. Parts of this production chain can use different technologies that are optimised for specific purposes and chaining different technologies and programming languages together, have sometimes been difficult. The technologies to support data science and machine learning processes are often built to support multiple languages by default and new solutions have been developed to create automated production chains by combining different functions, methods and code and to use metadata to control the execution.

Beyond support for multiple languages and metadata-controlled automation, these solutions are also created to scale efficiently. Solutions such as *Airflow* and *Kubeflow* allows for specific part of the production chain to be executed on specific, high-performance compute resources. This enables the sparse high-performance computation capabilities that might be available on-premises to be used only when it is necessary or even to scale-out specific processing on another environment.

Portability, replicatability and isolation of environments

The technological advancements in the form of *containers* have proven incredibly useful for cloud computing. More efficient use of compute resources, improving software portability and with the possibility of increasing security aspects are some of the benefits that the container technology has brought to cloud computing. Since container platforms can be setup on premises as well, this opens up great potential in supporting the requirements on flexible, secure and cost-efficient solutions for statistical production, even for organisations that aren't able to use cloud-computing solutions. Setting up specific it-services as well as whole environments (containing multiple it-services) can be automated using container technology.

Since the traditional way of setting up development test and production environments can be quite a complex task, it is not unusual that shortcuts are taken here and especially with solutions built by business where changes are sometimes deployed directly to production. The iterative form of the GSBPM also requires a flexible way of switching between building/developing, testing and production while maintaining control over what data was produced by what version of the production logic/code. Container technology also enables us to have multiple production environments running simultaneously which opens up the possibilities of trying out new production logic in parallel with ongoing production and comparing outputs from different production logic/code. To keep track of what service instances or environments are used for what, the container solutions should be context aware. Statistics Sweden has developed a separate system (UDB –translated to *Statistical program database*) to keep track of all *Statistical programs*, *Statistical program cycles, collection phases* etc and are currently

experimenting on using this information as metadata for our container platform to control such things as Kubernetes namespaces tied to *Statistical program cycles.*

If we return to the case above where shared services/platforms has been delivered by IT which allows statisticians to make larger changes to the design/production logic without new releases by IT. Here it will be difficult to develop new production logic without using real data, the development environment often must contain real data. Therefor the requirements on the development environments are usually the same as for the production environment. This suggests that from a GSBPM perspective, the container environments should use the context metadata to tag what services/environments is used for what but to use similar or the same templates for generating both environments to support development as well as production. The different tags (for production logic versions, development/test/production-status etc) can also be used by the container environment to fine tune compute resource management/prioritisation and to fine tune security policies.

Conclusion - Flexibility, quality *and* cost effectiveness

By building upon the technological advancements made as part of the data science, machine learning and cloud computing paradigm, we think that solutions could be built that deliver on all the crucial characteristics – *flexibility*, *quality* and *cost effectiveness*.

Using the same platforms for both traditional statistical production and for machine learning initiatives will help speed up adoption of machine learning in official statistics and using open-source solutions opens up the possibility for further collaboration within the statistical community while avoiding the risk of the lock-in-effect that cloud specific solutions can have.