A Unit-base derived from the SBR

Springboard to a Data Lake

Centraal Bureau voor de Statistiek

SN Strategic Agenda

# Be innovative!

Open up new data sources and use new methods

Make the innovation process more effective and efficient

# Towards a state-of-the-art data and information infrastructure

Implement a flexible, fit for purpose infrastructure

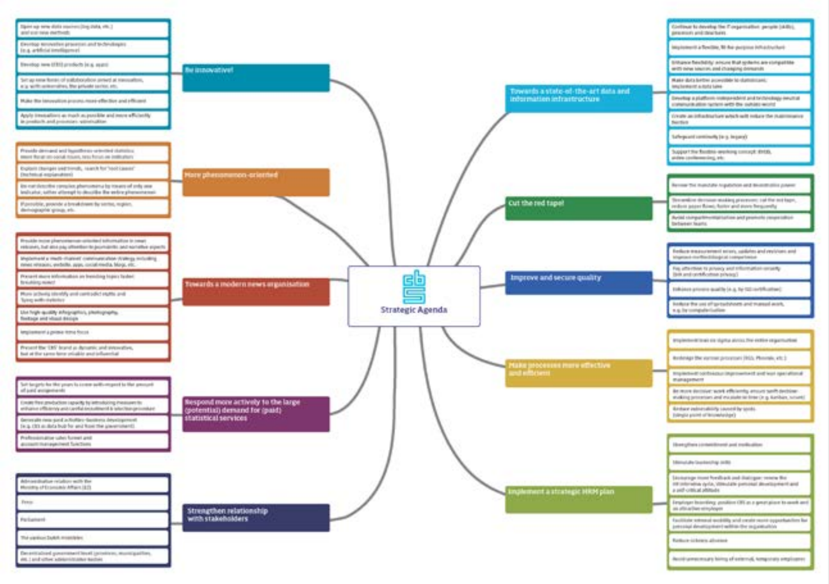
Make data better accessible to statisticians ; implement a data lake

# Make processes more effective and efficient

Reduce vulnerability caused by spoks

# Improve and secure quality

Reduce the use of spread sheets and manual work


Strategic Agenda
Be innovative!
More phenomenon-oriented
Towards a modern news organisation
Respond more actively to the large (potential) demand for (paid) statistical services
Strengthen relationship with stakeholders
Towards a state-of-the-art data and information infrastructure
Cut the red tape!
Improve and secure quality
Make processes more effective and efficient
Implement a strategic HRM plan

SN Strategic Agenda

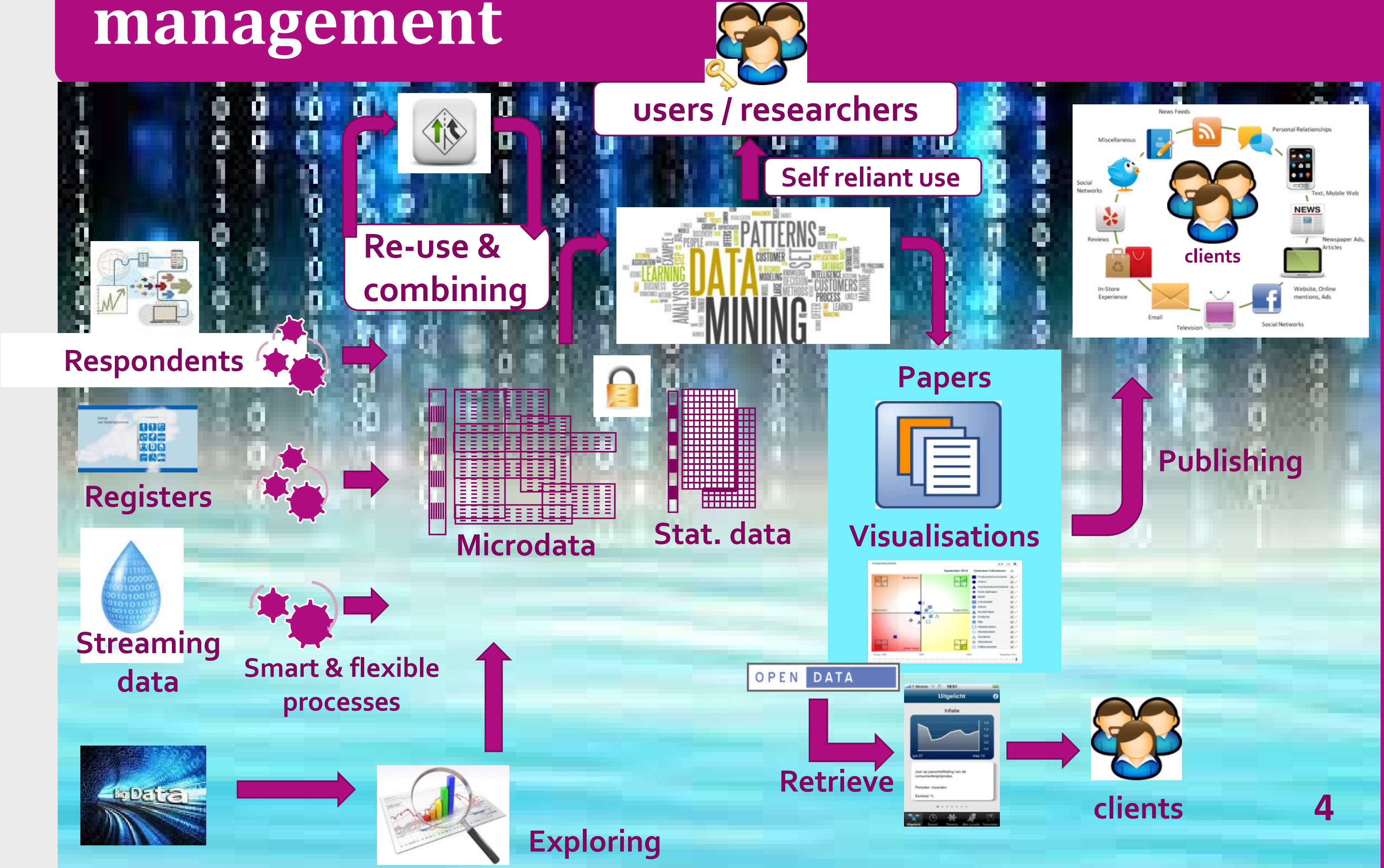
# Towards a state-of-the-art data and information infrastructure

Make data better accessible to statisticians; implement a data lake

CBS Data Lake definition:

"A concept to ensure that next to a ***decoupling*** *of input, processing and output, also the demand for* ***flexibility*** *and* ***coherence*** *is satisfied thereby guaranteeing that the information needs of the statistical producer and statistical user are fulfilled as* ***independently*** *as possible without the interference of methodology and IT support".*

™ For all aspects of future focussed data management



users / researchers

Self reliant use

Re-use & combining

Respondents

Registers

Streaming data

Smart & flexible processes

Microdata

Stat. data

Papers

Visualisations

Publishing

Exploring

Retrieve

clients

clients

Top 7 goals from end-user perspective

1. Enable **more phenomenon based output** (a phenomenon is a striking event that you want to explain)
2. Enable **more current and coherent statistics**
3. Stimulate the **reuse** of data
4. **Accelerate** the statistical **processes**
5. **Grow** and **stimulate** the **access** to a large number of **existing and new data sources**
6. **Provide faster response and output** to requests from external clients
7. **Accelerate the design process** around collecting and storing data

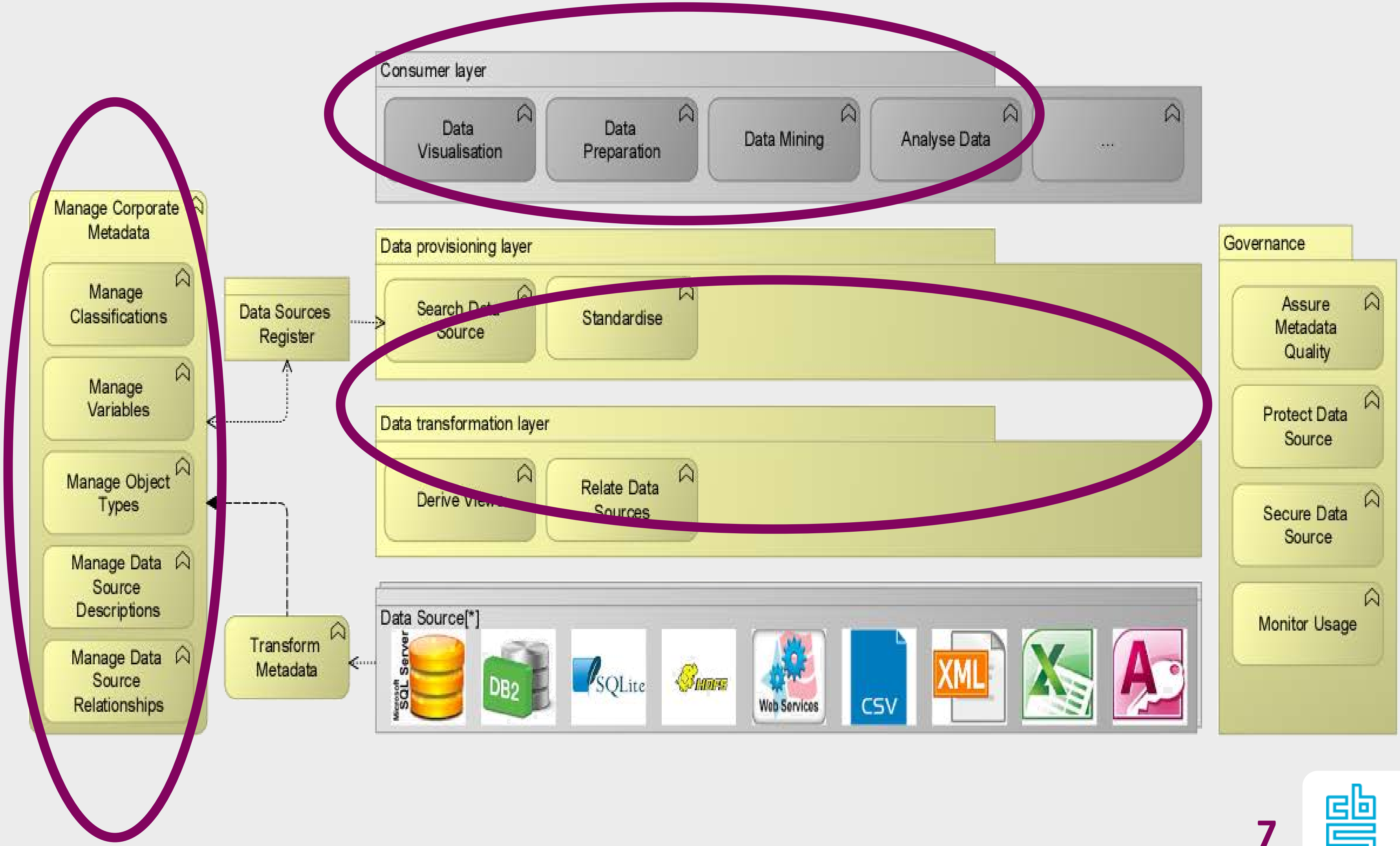
How to get there?
Enterprise Data Lake Project

- Project for a new architecture; **data driven**

- Focus on **end user goals**;
  - 5 Better accessibility of available datasets
  - 1 Dealing with many data sources, many formats
  - 7 Faster, phenomenon based reporting

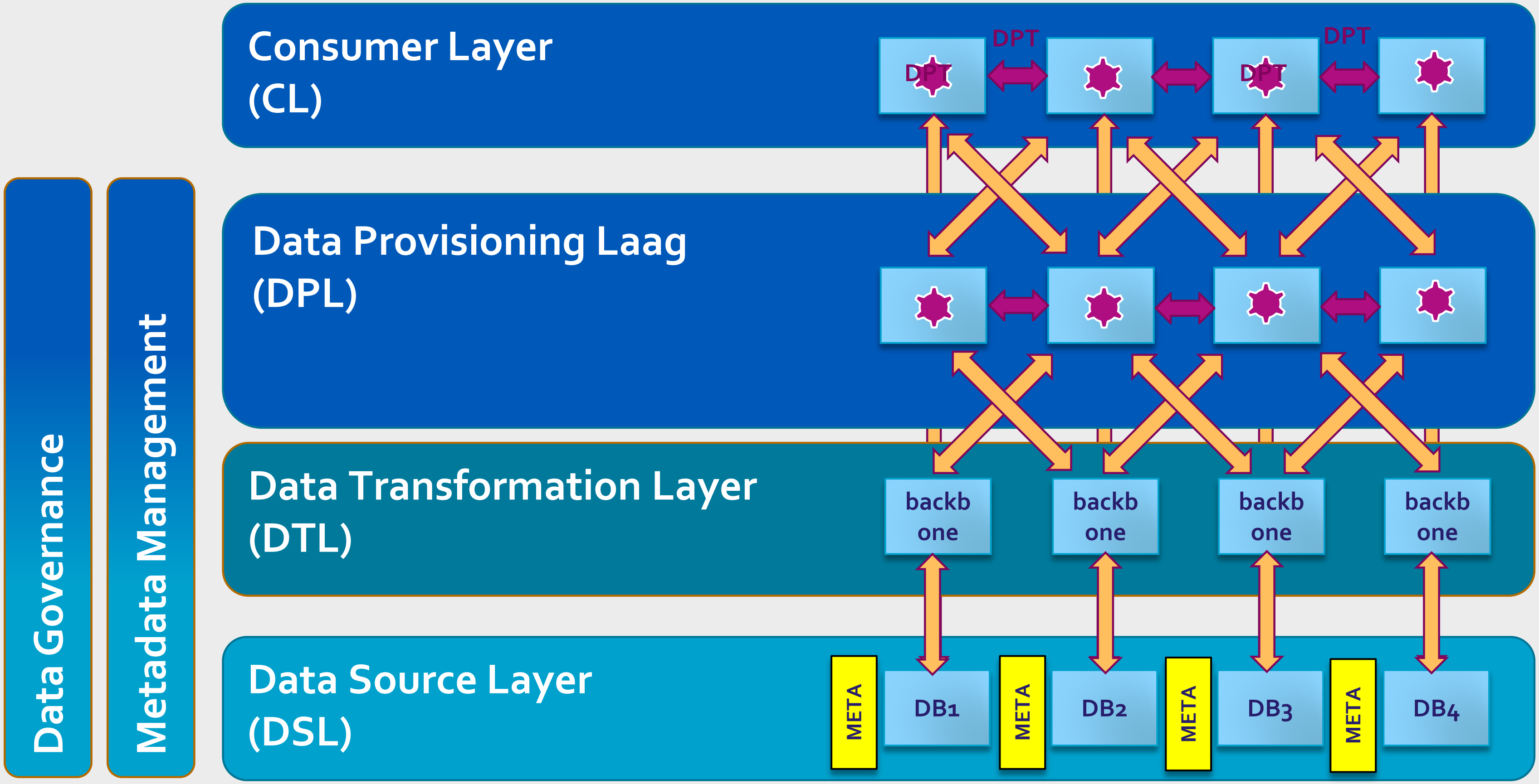
- Data Lake project consist of **three pillars**:
  - **Metadata** repository (technical & conceptual)
  - **Data Virtualisation** as technology to provide single data platform
  - User-friendly and self serving frontend by making use of **Data Preparation Tools** (DPT)

Data lake capabilities and main focus



Change to a Data Driven Architecture

Data consumers: custom fit, standard applications, scripts, batch etc.



Yeah great
but how about
the Statistical Business
Register ???

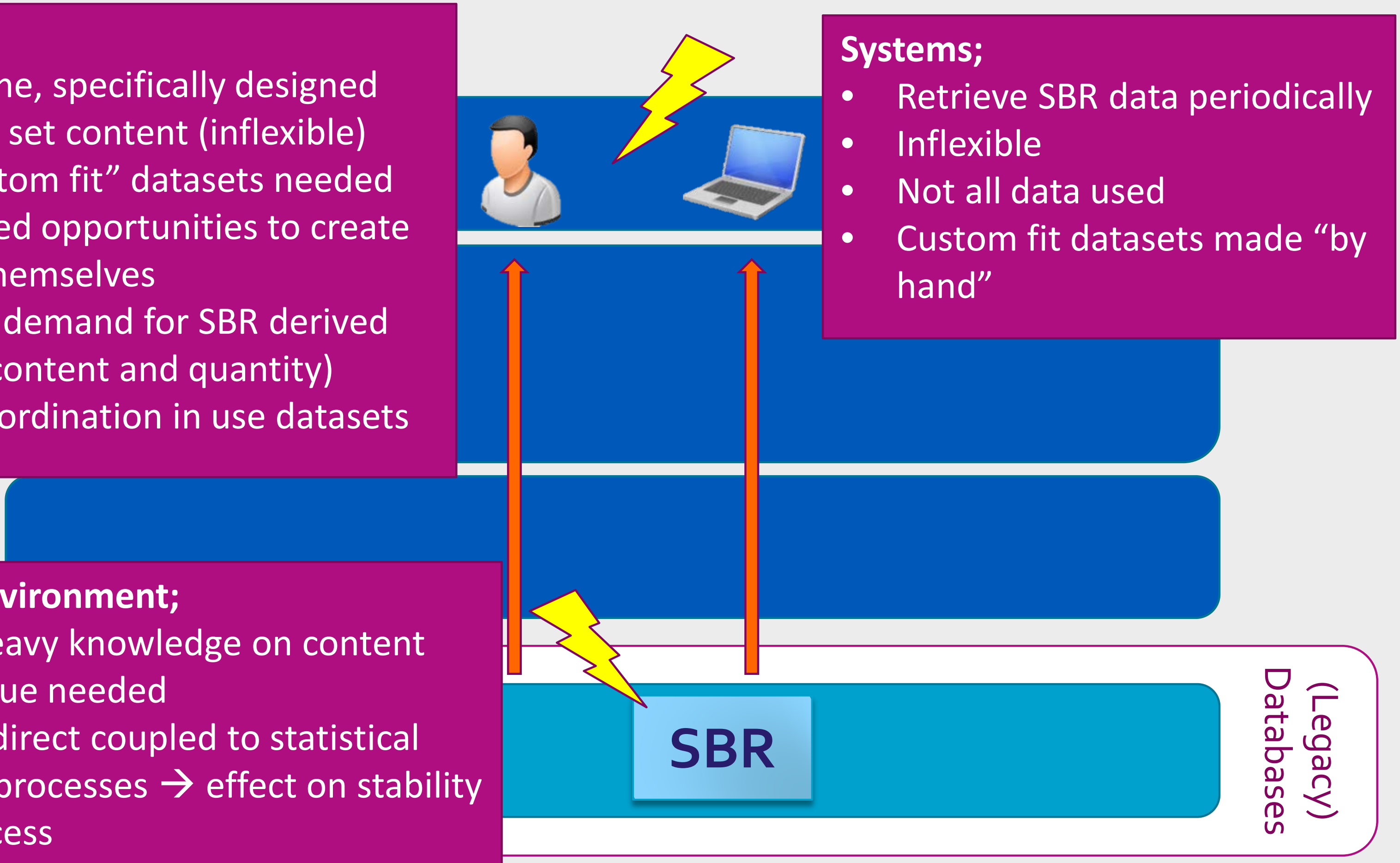
From…

Clients;

- At a set time, specifically designed and with a set content (inflexible)
- More “custom fit” datasets needed
- Have limited opportunities to create datasets themselves
- Increasing demand for SBR derived datasets (content and quantity)
- Limited coordination in use datasets

Systems;

- Retrieve SBR data periodically
- Inflexible
- Not all data used
- Custom fit datasets made “by hand”



SBR Process-environment;

- Complex, heavy knowledge on content **and** technique needed
- Technically direct coupled to statistical production processes → effect on stability of total process
- Not “in rest” → **Live Register**
- Snapshots and frozen frames in same system and from same system to clients

To:

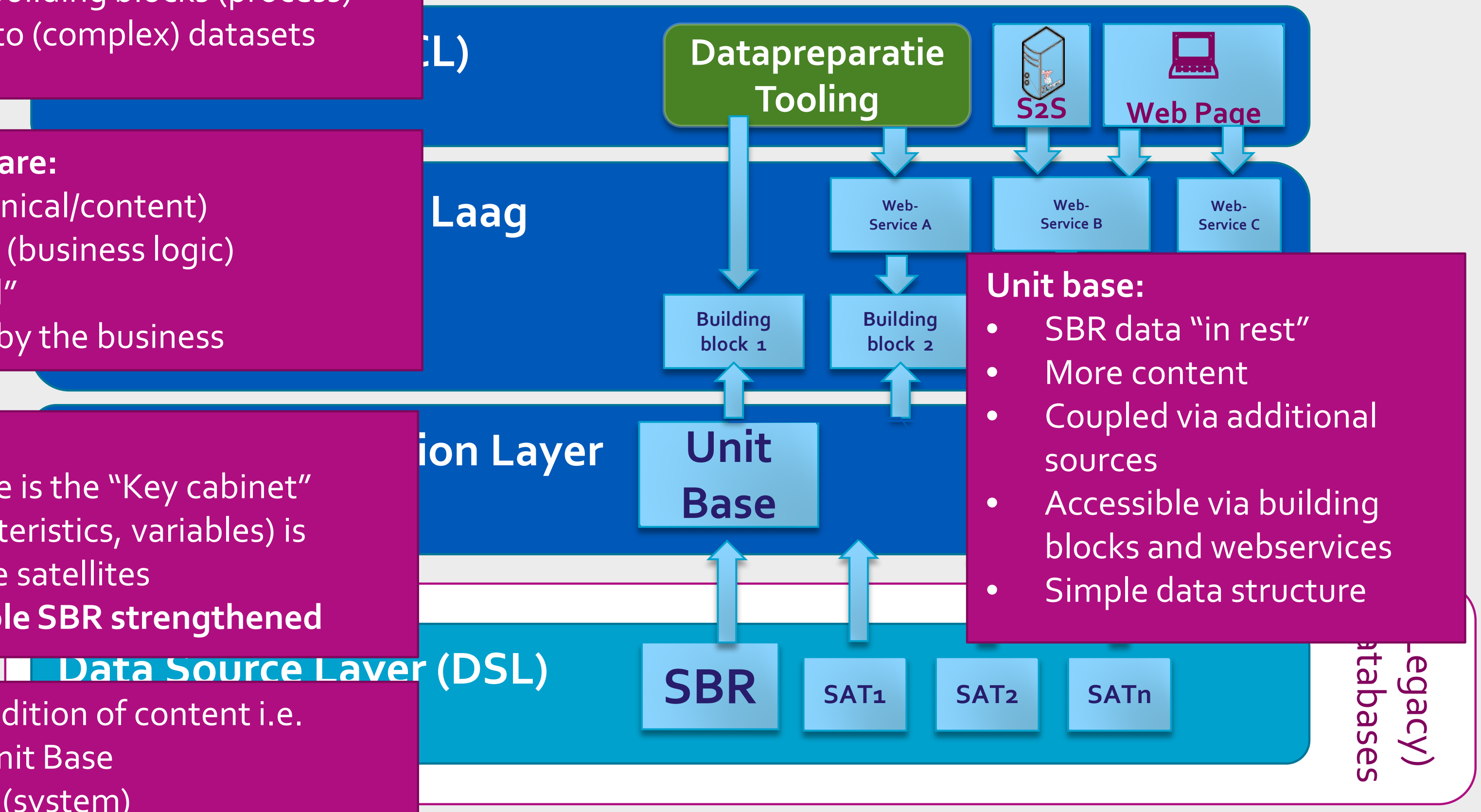
- Systems coupled via webservices
- Data “on demand”
- Webservices easy adjustable and expendable

Data preparation tooling:

- Easy use of building blocks (process)
- Easy access to (complex) datasets

Building blocks are:

- Simple (technical/content)
- Coordinated (business logic)
- “On demand”
- Expandable by the business



Unit base:

- SBR data “in rest”
- More content
- Coupled via additional sources
- Accessible via building blocks and webservices
- Simple data structure

DTL:

- The Unit base is the “Key cabinet”
- Data (characteristics, variables) is added via the satellites
- **Backbone role SBR strengthened**

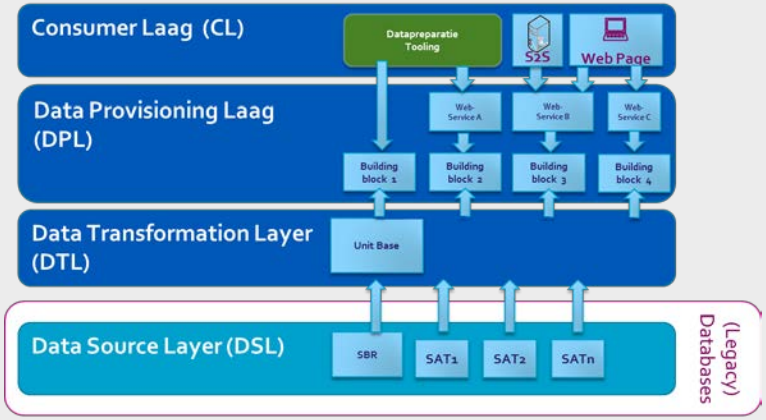
- Unlimited addition of content i.e. linkable to Unit Base
- Outside SBR (system)
- **SBR as a core of SU, not complicated by surplus data**

Results

- By realizing the Unit base established 0-version of data driven architecture (concepts) .
- Users have **on-demand and easy** access to a wide and expandable set of SBR coupled data.
- Building blocks are **adjustable** and **expandable** without IT interference (webservices).
- Increased use of SBR (coupled) data
- System in rest; **decoupling** of **Statistical Business Register** processes and other sources.
- Unlimited **addition of content** (characteristics, variables) that can be linked to the Unit Base
- **SBR as a core of SU, not complicated by surplus data → true backbone role**

Unit base as a springboard…

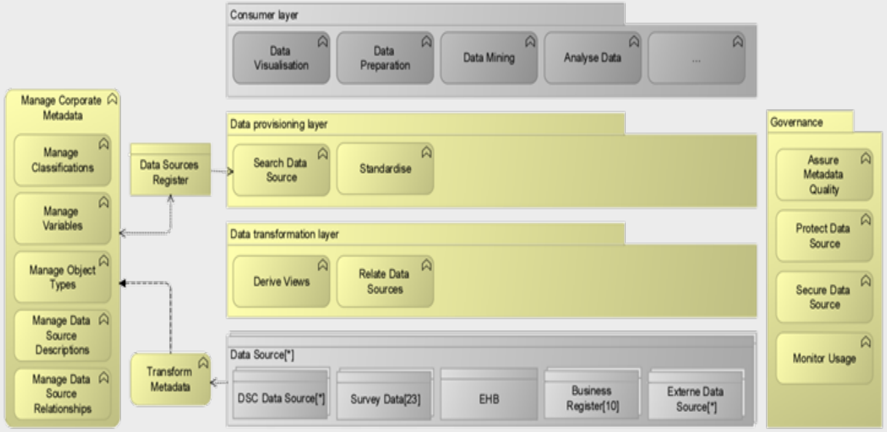
Unit base



- Scope SBR department
- Only SQL Server sources
- Physical datastorage
- Implicit metadata management

With Unit base proven that concept of Data lake works

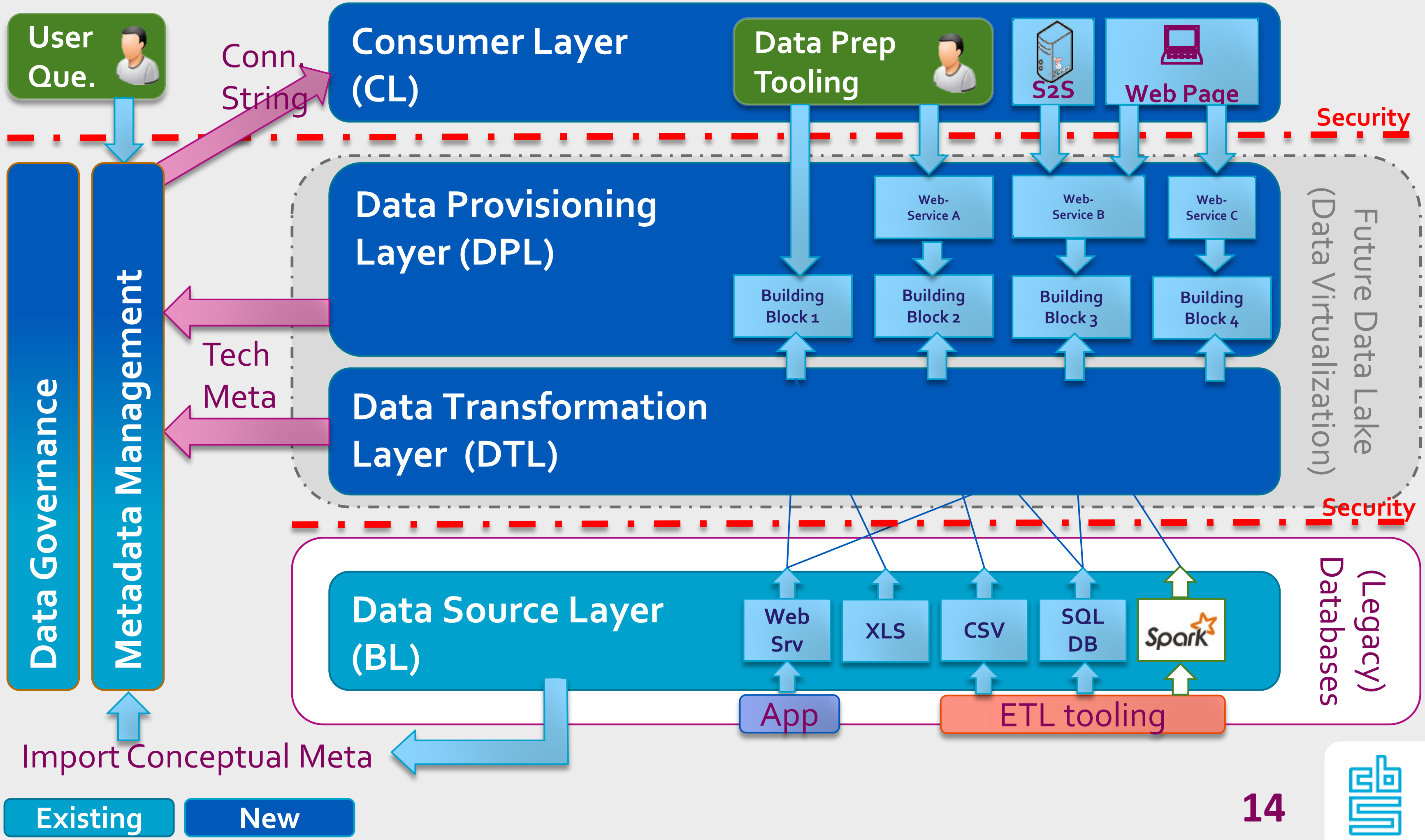
Data lake



- Scope Statistics Netherlands
- All possible sources
- Virtual data
- Explicit metadata management
- Extensive testing of commercially available tools

Buildingblocks from Unit Base can be re-used

The new architecture



supply | demand

Thank You!

Contact information:
Irene Salemink
ISLK@CBS.nl