Dutch Enterprise Data Lake
Fishing in clear water
Irene Salemink
NSI 2.0 National Statistical Systems

• We make data available in an integrated, flexible and controlled manner

• We offer a platform for collaboration between authorities
Relations

Data sources
- Police
- Local health authorities
- Chamber of commerce
- Population register
- Businesses
- Social benefits register
- General public
- Tax authority
- Education
- Land registry

Users
- Ministries, Policy analysts, Central bank
- Universities
- Parliament
- Local government
- Private sector
- Education
- Health care
- The public
- Media
- International

Collaboration
Phenomena

**Factsheet**

**The Netherlands:**

- **Cycling country**
- **2.9 km average daily distance cycled by the Dutch**
- **25% of home-work commuting is done by bike**
- **1.4 million electric bikes in 2014**

**Provinces with the most and least metres of cycle paths per ha of land area, 2015**

- **35,000 km total length of Dutch cycle paths**
- **240,000 Dutch practises cycling as a sport**

- **1,000 km average distance cycled per person per year**
- **2,000 km average distance cycled per teenager per year**

- **558,000 Dutch victims of bike theft**

**Traffic deaths, 1996-2014**

- **5%**
- **4%**
- **15%**
- **41%**
- **20%**


Publication date: 3 July 2015
Great Ambitions...

Progressing towards a data-driven society
... Sustainable Development Goals

- Economy
- Education
- Energy
- Environment
- Finance
- Fire & Emergency Response
- Governance
- Health
- Recreation
- Safety
- Shelter
- Solid Waste
- Telecommunications
- Transportation
- Urban Planning
- Wastewater
- Water & Sanitation
...but also great Challenges
Connecting data... Data lake?

- Always recent data
- Distributed Data
- Sensitive Data

Information security and Access management are very Important!
Stakeholders

End-users
- Internal CBS
- External
  - Data access, Re-use of data and designs
  - Coupling & Combining
  - Efficiency & flexibility

Source owners
- What happens with the data?
- Authorisation & Security

Sponsors
- Internal (CIO, Controller) → Business Case?
  - External (Ministries, Governmental bodies, private parties)

(security)custodians and other environment
- Legal mandate, ethical concerns
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 as possible without the interference of methodology and IT support”.
A Data lake is a.....?

TechTarget: A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed....each data element in a lake is assigned a unique identifier and has extended metadata tags....when a question arises the data lake can be queried for relevant data.

CBS Data lake: confined to statistical data. These data describe economic and social phenomena and have therefore a structure concerning the content and a semantic meaning. It is a logical data warehouse, integrating data sources in real time, without data duplication, regardless structure, technology or location.

Gartner: A data lake is a collection of storage instances of various data assets additional to the originating data sources...in a near-exact/exact copy of the source format. The purpose is to present an unrefined view of data to only the most skilled analysts to help them explore their data refinement and analyse techniques independent of any of the system-of-record compromises that may exist in a traditional analytic data store.
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 **re-use** 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 for a new architecture; data oriented
- Focus on end user goals;
- Better accessibility of available datasets
- Dealing with many data sources, many formats
- Faster, phenomenon based reporting

- Data Lake project consist of three pillars:
  - Metadata repository (technical & conceptual)
  - Data Virtualisation as technology to provide single data platform
  - Security and Authorisation to prevent data sets from unauthorized use
BA...... from process oriented
... to a data oriented approach
Key Capabilities

Ability to:

- Discover, access and understand
- Load, store, model, retrieve
- Transform, harmonize, integrate
- Access, derive, catalogue
- Use (prepare, visualise, analyse...)
- Manage as an asset
- Secure
Capability Groups

- **Consumer Layer (CL)**
  - Data Preparation
  - Data Visualization
  - Data Analyses
  - Selfservice Reporting
  - Search & Explore
  - Data Reading/Scorecarding
  - Messaging
  - Data Mining

- **Data Provisioning Layer (DPL)**
  - Data Access
  - Data Hub
  - Data Aggregation
  - Derive Views
  - Data Catalog

- **Data Transformation Layer (DTL)**
  - Data Harmonization
  - Data Transformation
  - Data Enrichment
  - Data Validation
  - Model Data Source

- **Data Source Layer (BL)**
  - Data Storage
  - Data Load
  - Data Access
  - Data Deletion
  - Model Data Source
  - Data Extraction
  - Ingest data
Key Building Blocks

- Metadata Model
- Semantic Technology
- Data Virtualisation
- Big Data Platform
- Self-Service BI / workflow orchestration
What does the Data lake offer?

- **Metadatamodel** that describes statistical data in a formal and exact way to map any statistical dataset to model represented as a graph and use meta to find data (including ranking)
- **Metadata management system** to manage and harvest technical & conceptual metadata
- **Data Governance and Security model** for managing and securing (shared) virtual datasets
- **Virtualisation** to decouple Data Source Layer from Consumer Layer and create virtual datasets / virtual views in order to retrieve, combine and process data without moving or copying data
- **Front end** that is user-friendly and self-supporting by making use of a Data Preparation Tool
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

**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:

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

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

Unit base:
- Systems coupled via web services
- Data “on demand”
- Webservices easy adjustable and expendable

- SBR data “in rest”
- More content
- Coupled via additional sources
- Accessible via building blocks and web services
- Simple data structure
Statistics Netherlands: national data hub

- **CL** = Consumer Layer
- **DPL** = Data Provisioning Layer
- **DTL** = Data Transformation Layer
- **DSL** = Data Source Layer
- **P** = Data Prep
- **V** = Data Visualization
- **A** = Data Analytics

**User Que.**

**Data Governance**

**Metadata Management**

**Tech Meta**

**Zone CBS**
- Web-Service A
- Building Block 1
- Building Block 2

**Zone CBS**
- Web-Service D
- Building Block 7
- Building Block 8

**Zone Tax**
- Web-Service B
- Building Block 3
- Building Block 4

**Zone Tax**
- Web-Service C
- Building Block 5
- Building Block 6

**StatLine**

**Secured VPN**

**Supp | Demand**

**Security**

**Data Sources**
Recommendations

- Check whether your **strategy** is in line with your plans (v.v.)
- Start **experimenting with Data Virtualization** in an early stage
- Build a culture that **embraces change** and **communicate** your plans as often as possible
Thank You!

Contact information:
Irene Salemink
ISLK@CBS.nl
Data protection

- Privacy is guaranteed (confidentiality required by law)
- All staff are required to sign declaration of secrecy
- Data on individual persons are immediately separated from names and addresses
- Under the law, data may only be used for statistical purposes
- No other institution may claim access to data collected by
Data Virtualisation in a nutshell

Connect
- Connect disparate data from any CBS source (DSC, Big Data, Cloud, Filesystem) or location

Combine
- Define (statistic) data transformations and combinations that meet the business needs.

Consume
- Deliver data services in real-time to the CBS data consuming platforms or tools.
What do we want to achieve with the Data Lake

- **reduc**
  - Cost data-access
  - Time to Market
  - Statistical Risk

- **stimulate**
  - Growth
  - Re-use
# Data Lake project – work in progress

<table>
<thead>
<tr>
<th>Status</th>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finished</td>
<td>4-layers Data Architecture</td>
<td>Possibility to decouple Data Source Layer from Consumer Layer and create virtual datasets / virtual views. Web Service interface implemented for business register EHB project demonstrated that architecture delivers benefits</td>
</tr>
<tr>
<td>Finished</td>
<td>Metadata Model</td>
<td>Develop Model that describes statistical data in formal and exact way. In theory it is possible to map any statistical dataset to model represented as a graph and use meta to find data (including ranking)</td>
</tr>
<tr>
<td>Finished</td>
<td>PoC Data Virtual</td>
<td>Successfully connected Denodo to Documentum Database (DSC) / improved query possibility &amp; performance boost</td>
</tr>
<tr>
<td>In Progress</td>
<td>PoC Metadata</td>
<td>Implement metadata model in PoolParty semantic web platform, harvest technical &amp; conceptual metadata and provide URL to DV platform</td>
</tr>
<tr>
<td>In progress</td>
<td>Connect Data Sources</td>
<td>Expand number of Data Sources to improve usability of test platform. Perform stress tests</td>
</tr>
<tr>
<td>Scope defined</td>
<td>PoC Multi-Zone DV</td>
<td>Use Data Lake as a research platform for distributed data. Implement secure infrastructure</td>
</tr>
<tr>
<td></td>
<td>Data Governance and Security</td>
<td>Define Data Governance for managing and securing virtual datasets</td>
</tr>
</tbody>
</table>