Virtual Ensemble Modeling
The road to the logical virtual model
Antoine Stelma

• Lead architect Connected Data Group
• Co Founder Connected Data Group & Connected Data Academy

• Data Architect +20 years
• Data Vault certified in 2006
• First Data Vault certified customer site in 2008
• Background in DWH automation, Data Modelling, Cloud & Data Virtualization specialist
• Spare time......
Challenges in the great data era

My Challenges

• Modeling
  • Data Vault modeling as the great inspiration
  • Not all data needs to be stored from a time historical perspective
  • ELM got introduced by GA
  • Can a Virtual Ensemble Logical Model work?
  • Teamed up with Rick van der Lans to combine Data Virtualization and Data Vault (SuperNova)

• Storage
  • Moving from on-premise DWH to cloud Data Platforms
  • Separation of the data model and physical storage
  • Integration can be done virtual
  • Got addicted to the concept of Data Virtualization

• Customer case and next steps
What is Data Virtualization?

- **Virtual Modeling**
  - Separates compute from storage
  - Allows virtual modeling based on all types of data sources
  - Data model and data storage are no longer a package deal
  - Flexible way to adjust virtual data models
  - Harvesting meta data from data sources
  - Data modeling is essential

- **Security and access**
  - Full monitoring on performance, access by data users
  - Provides a central semantical access layer for all data users
  - Centralized security model, based on Active Directory

- **Data Virtualization extra’s**
  - Intelligent query engine to optimize queries towards data sources
  - Lineage up- and downstream
  - Horizontally scalable for High Availability use cases
  - Delivers end points and a data catalogue for data users
Data Virtualization: key capabilities

Data Abstraction:
- specify, share and personalize for multiple personas

Technology Abstraction:
- understand the data and specification over tools & technology

Flexible Data Architecture:
- collect and connect to any data source, for any use case

Data Management & Governance:
- data quality, data engineering, (meta/master) data life cycle, security
Concept of SuperNova

Written by Rick van der Lans in cooperation with Connected Data Group

• Virtual joining of Hub and Satellites to one virtual view including time historical data.
• Virtual joining of Links and Satellites to one virtual view including time historical data.

• Pros
  • Can be completely generated based on a proper Data Vault
  • The physical data vault enforces the logic
  • Easy start point for virtual star schemes or other data products
  • By separating the storage from the model, changes are much easier to apply

• Cons
  • Performance of Data Vault can be an issue
  • Link Satellites......
  • Still not working towards a logical model

• Download the whitepaper here
Lineage of virtual SuperNova

Virtual introspection of the physical Data Vault

Virtual metadata driven harmonisation objects

Virtual generated SuperNova Model

Virtual build Sales table to Publish as data or webservices

Security policies and audit use of data

Virtual objects generated based on metadata
Next Step: Virtual ELM

Housing corporation customer case

• Customer has multiple data sources and wants to add external data to the virtual logical DWH
• Customer bought Tibco Data Virtualization and adopted our methodology
• Customer was trained and familiar with the Data Vault concepts by GA.
• Requirements for the Virtual ELM model:
  • Use the logic of current Data Vault Data Warehouse, but not as a source
  • Primarily interested in latest version of the data, but needs possibilities to store and use historized data without building extra logic
  • Add extra, external, data to the model from external information suppliers
  • Create virtual Star Schemas
  • The models in Data Virtualization must be fully logical and translations from source to model must be metadata driven
  • Must be modular and easy to scale out
Virtual Publication Layer

- **Real Estate**
- **Agreements**
- **Relations**

- **Revenue Rental Agreements**
  - Agreements 75%
  - Real Estate 15%
  - Relations 10%

- **Rental Real Estate Efficiency**
  - Agreements 100%
  - Real Estate 100%
  - Relations 90%

Virtual Business Layer

- **Source A**
- **Source B**
- **External source**

Virtual Source Layer

- **CBC**
- **NBR**
- **Cache**
- **History storage**
Our approach (1/2)

- Identify the heart beat of the organization (Step 1: Identify)
  - Discover the three most important entities that run the business
  - This helps to set up a solid base for the virtual ELM
  - Create a logical model on the first three entities
    - Create the mapping between the logical model
    - Create the metadata to translate the source entities to the virtual ELM entities.
    - We used the knowledge and the current Data Vault Data warehouse to identify the CBC and the NBR

- Started with the most important virtual ELM: Real Estate (Step 2: Base ELM Model)
  - After introspection of the ERP data source, we generated based on the design and metadata the harmonization layer.
  - This creates a solid starting point for building the first virtual ELM
  - By dividing the ELM in business-friendly items, we sliced the ELM virtual in reusable views.
  - These views can be used as a stand-alone dataset or as a part of the virtual ensemble, call “ELM-real estate”
  - Dividing the ELM also helps out with part of the data that is static versus changing frequently
  - Data per view can be cache of store permanently to keep track of changes through time
  - The ELM-real estate virtual Ensemble Model is represented as on virtual data
  - The smart engine of Tibco Data Virtualization will generate based on this logic the optimal query to the source systems
Our approach (2/2)

• Deal with relations (Step 3: Relation Model)
  • A relation (link) was first built as a separate view, like Data Vault (old habits...)
  • A link is now a part of the virtual ELM, by using the business keys
  • Every virtual ELM has its own business keys to connect to other virtual ELM.
  • If information about a relation needs to be store (minimal) this will be done in the connecting virtual ELM.
  • So the virtual ELM Agreements contains the information about the relation with the ELM Real Estate and so on
  • All logic is virtual, so Data Virtualization will push down to the data source to solve a request.
  • On top of the Virtual ELM’s a Star Schema is built for end-users

• From end user to source and back (Step 4: End User Model)
  • End users queries virtual star schema
  • Data Virtualisation receives the request and based on the virtual ELM model, it will write a query to the relevant source systems making sure that only the relevant data is queried from the data sources
  • If necessary, Data Virtualization can store data permanently for historization or cache result to support performance
• Concepts of Data Vault for virtual Ensemble Logical Modeling are key
• By using Data Virtualization, combine Data Vault Data warehouse with other sources
• The ELM concept is great way to work top-down
• Metadata should be used to generate and prepare the base of the virtual ELM
• By using virtual logic models end users can be a part of the journey
• Just to have only the virtual ELM for Real Estates is already big step for end users
  • All data integrated and managed into one virtual view for reporting and analysis
• The need for Master- and Metadata management is a spin off for this project

Next steps
• Store data with Azure Table Storage and/or Azure Synapse to optimize performance while virtual modeling will be done in Data Virtualization
• Research in motion: how to deal with meaning less keys to automatically link virtual ELM
• Writing down and working on the concept with GA
Data Virtualization & ELM

Ensemble Logical Modeling

Logical business model

- Leveraged for all logical modeling needs including the data warehouse, big data/lake, streaming, master data management (MDM), Blockchain and operational integration initiatives.
connected data group

Data & Analytics Services

Antoine Stelma
T. +31 6 287 682 26
E. antoine.stelma@connecteddatagroup.nl

Erik Fransen
T. +31 6 159 444 76
E. erik.fransen@connecteddatagroup.nl