Informatica

1 June 2022

Architect Workshop

Data Lineage - Do you navigate your data or still ask for directions?

VGH
fair versichert

Housekeeping

- This session will be recorded
- Please feel free to ask any question via the Q&A option (not via chat) – they will be answered asap, some questions may be taken at the end of the session
- Please interact with us via the poll questions asked during the presentation
- End time is 16:30 CET

Informatica

Agenda

1. Types of data lineage
2. Automation best practices, pitfalls and limitations
3. Customer case: VGH Versicherungen
4. Q&A

Informatica

Types of data lineage

Tobias Rebele

Data Governance Domain Expert - DACH

Enterprise Data Lineage

Enterprise Data Lineage

Business Lineage

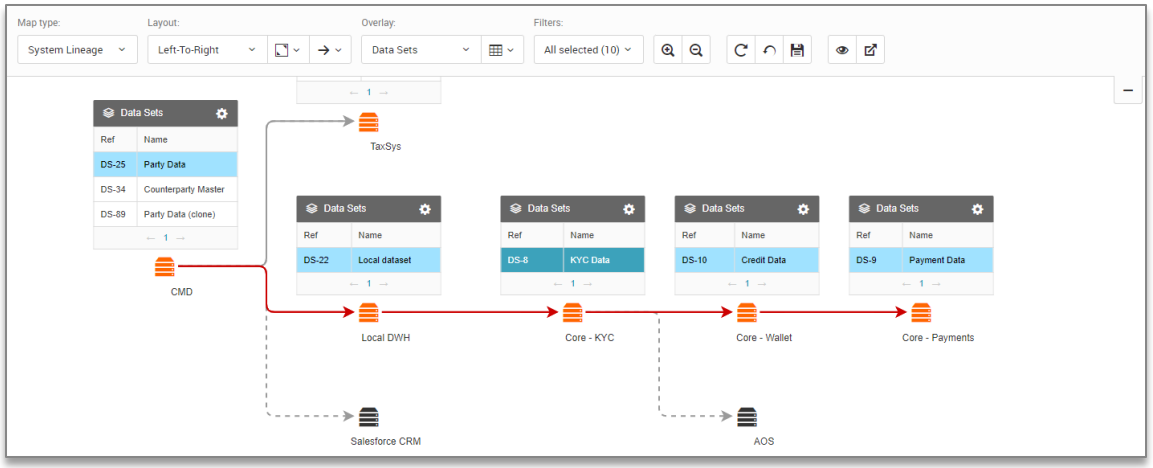
- Business-friendly representation
- Business-relevant data elements and their business context
- User-defined flows

Technical

Vertical

Enterprise Data Lineage

Business Lineage



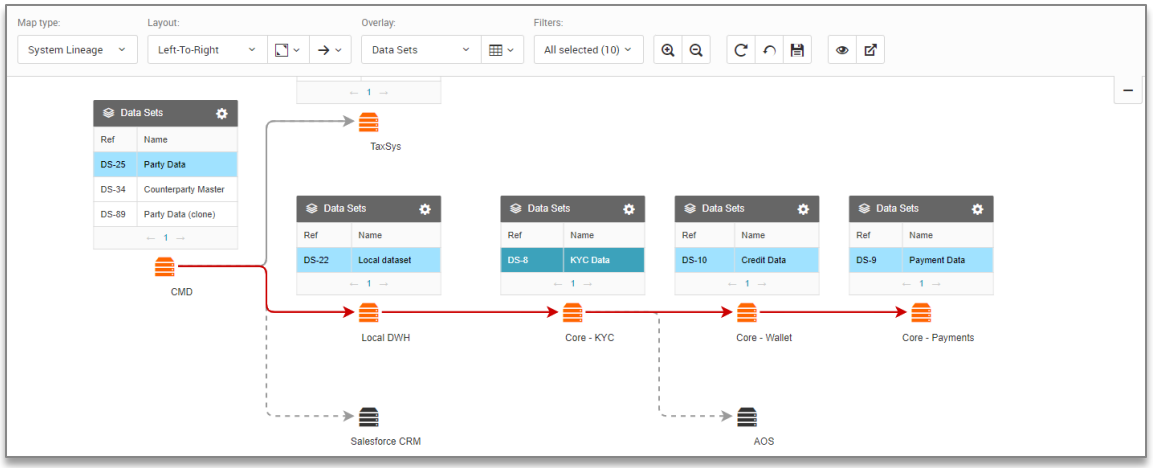
Technical Lineage

- Automated extraction from complex enterprise systems
- Automated parsing of code from stored procedures in databases and multi-vendor ETL tools – both, static and dynamic code
- Complete visibility into procedure calls with parameter tracking and dynamic SQL generation based on parameter values

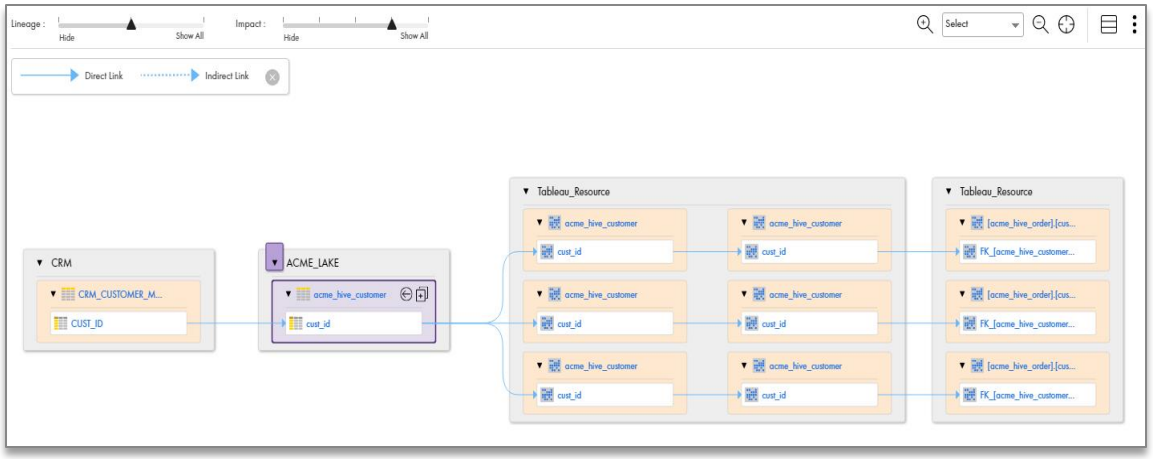
Vertical

Enterprise Data Lineage

Business Lineage



Technical Lineage



Vertical Lineage

Lineage approaches

Technical Lineage

Where? How?

- Pros:
  - Faithful representation of reality
  - Can be automated
  - Can be kept up-to-date, if automated
- Cons:
  - Very hard to reverse engineer legacy
  - Overwhelming for business users
  - Automation doesn't capture the business semantics

Business Lineage

When? Why?

- Pros:
  - Easily consumable by business users
  - Can be easily linked to business semantics
  - Can be built and maintained by users
- Cons:
  - Requires more business user engagement
  - It is an abstracted view
  - Can be incorrect / out of date

Informatica

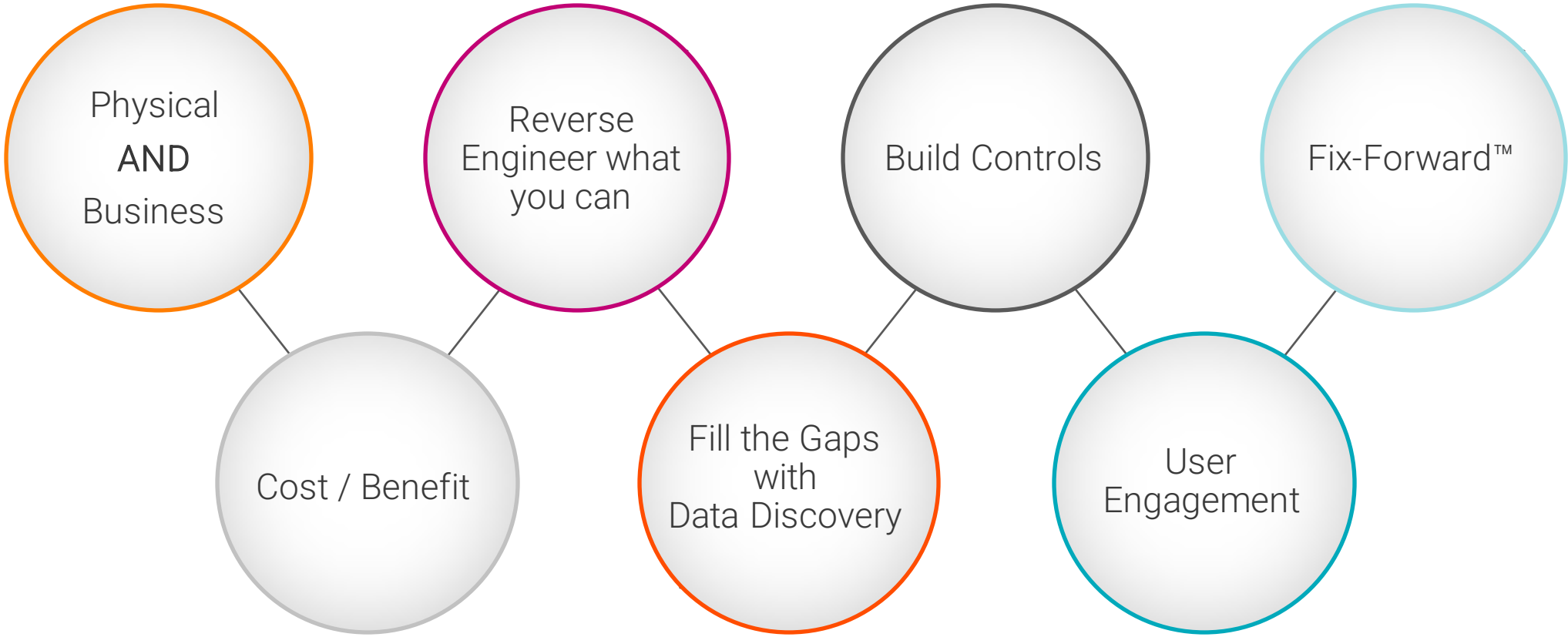
Automation best practices, pitfalls and limitations

Remy van der Kleij

Solution Architect – Benelux & Nordics

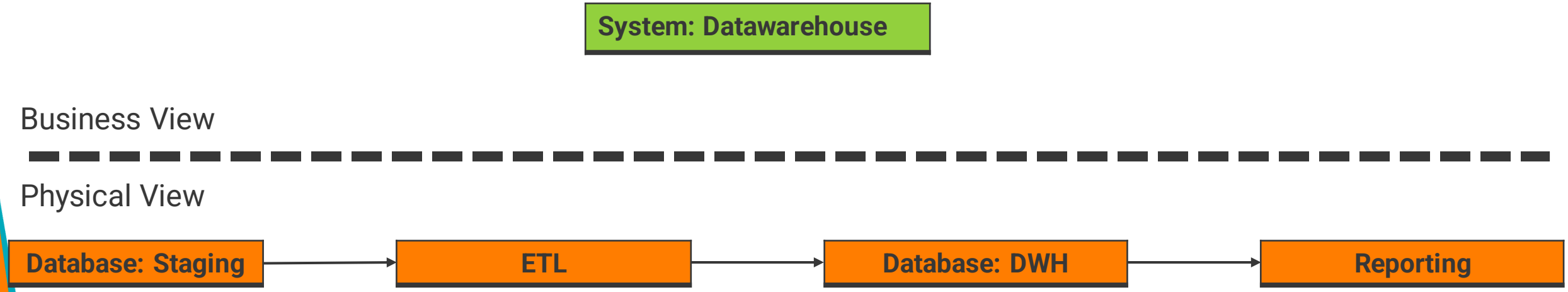
Best Practices for Data Lineage

Focus on transparency, provenance and clarity



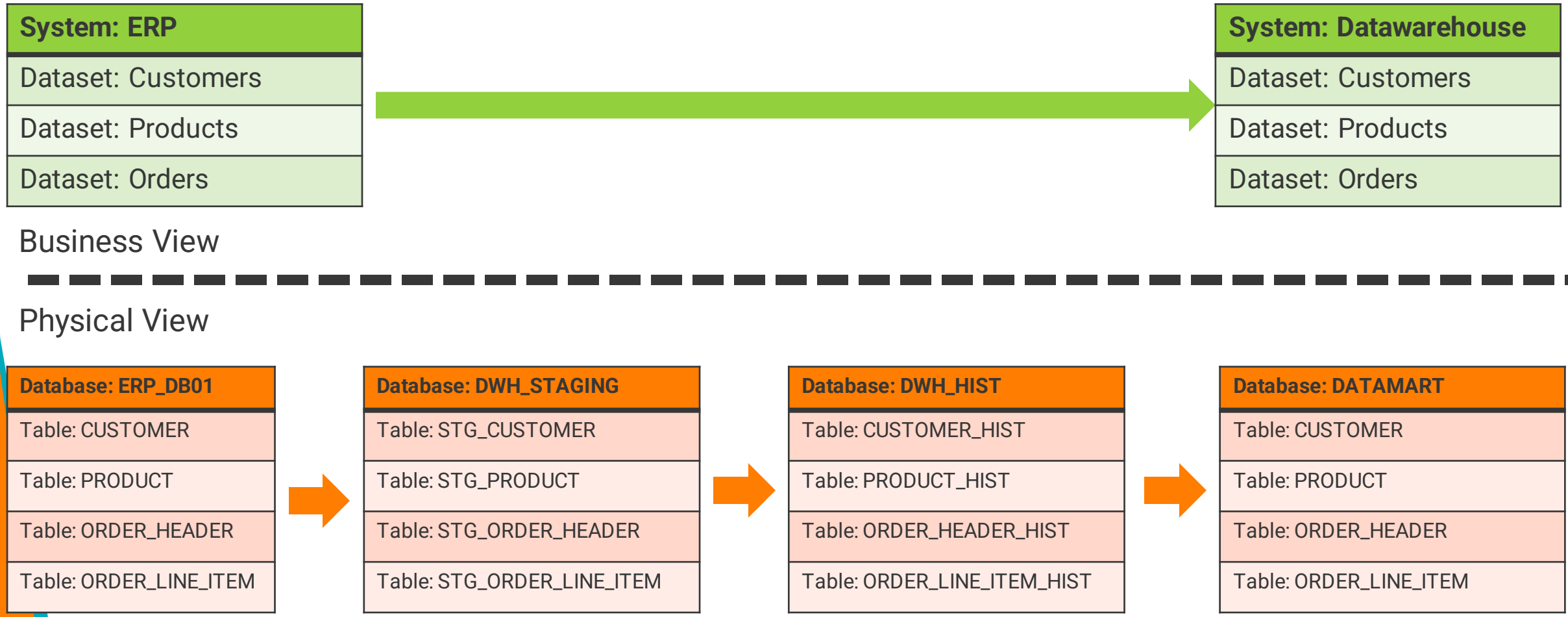
Informatica

Systems and Resources



Informatica

Staging / Technical tables



Informatica

Key Value Pairs - structure

Structure

Content (example)

Dataset: Customer
Identifier
Full Name
Telephone Number
Email Address

Identifier	Full Name	Telephone Number	Email Address
1	ACME Corporation	+1(123)456789	info@acme.com

Business View

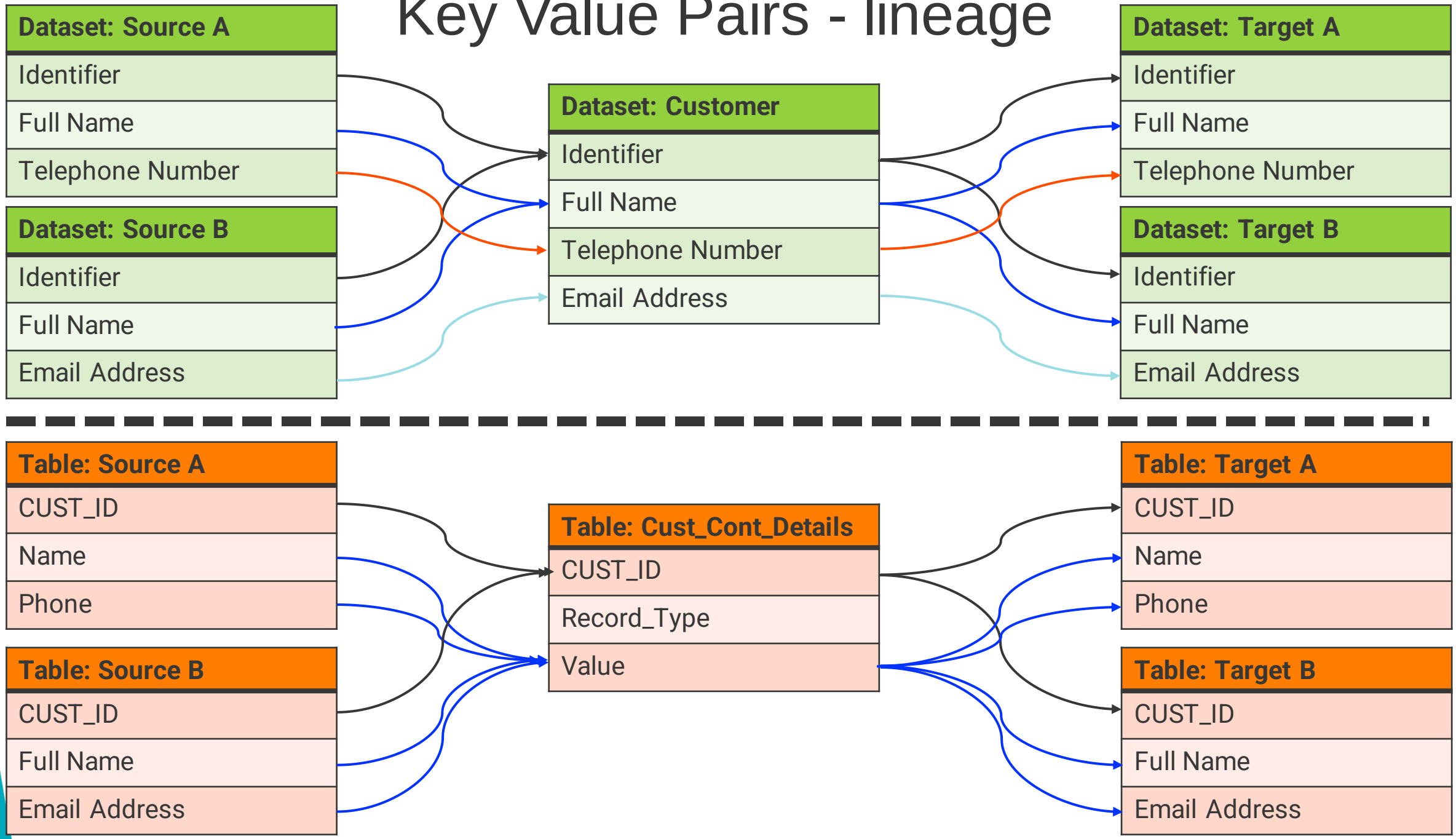
Physical View

Table: Cust_Cont_Details
CUST_ID
Record_Type
Value

CUST_ID	Record_Type	Value
1	FullName	ACME Corporation
1	Phone	+1(123)456789
1	Email	info@acme.com

Informatica

Key Value Pairs - lineage



Automation Pitfalls & Limitations

What to avoid / reconsider?

- Don’t expect business lineage to be a simple ‘summary’ of physical lineage
  - Physical data structures are often not understandable for business users
  - Business requirements may require splitting physical assets into multiple business assets
- Balance use of dynamic parameter-driven frameworks versus ‘traditional’ development
  - Complex to reverse-engineer frameworks because metadata is dynamic and only available at runtime.
  - Include lineage support in data integration design requirements → “Fix Forward”
    - I.e. require frameworks and integration tools to support data lineage
- Don’t ignore chances to adjust existing technology to enable/simplify lineage
  - Add query logging to existing frameworks so lineage can be derived from regular SQL statements
  - Organize ETL parameter files (content, location, naming) to simplify lineage configuration

Customer Use Case

VGH
fair versichert

Informatica

Lineage & Impact Analysis with EDC @ VGH Insurances

Taming the Data Monster (with) Informatica EDC

Bernhard Link | VE-WP – RPA Manufactory | May 2022

VGH

BI@VGH: Metadata driven³

Highly standardized architecture for VGH's Enterprise Data Warehouse

- Based on highly **parametrized** and **object oriented** ETL architecture ➔ lineage challenges

Our three dimensions of metadata:

1. ETL architecture fully controlled by **metadata** (which is also used for reporting)
2. Selfbuilt application „BI Factory" as automization framework for **metadata** based design and generation of all relevant BI objects (Backend: DB, ETL, ParmFiles; Frontend: models, cubes; business logic excepted)
3. EDC as new **metadata** dimension and hub for various usecases

Not loaded to EDC

Metadata framework set up to generate EDC objects, e.g.

- REST commands for resource management (insert, update, delete)
- REST commands for **connection assignment**
- CSV data for bulk import of custom attributes
- CSV data for bulk import of **custom lineage**

VGH

Prio 1 Usecases EDC@VGH (1)

Support GDPR Compliance

Helping to fulfill requirements of GDPR by identifying (and protecting) personal identifying information (PII) and by collecting this metadata for our BI Factory ➔ Data Domain Discovery & REST API, EDC Security

Helping EDC users to answer questions like

- „Which path follows this PII attribute, that I want to anonymize, in the whole process chain?“ (BI developers, Data Protection Officer)

➔ Data Lineage & **Impact Analysis**

Prio 1 Usecases EDC@VGH (2)

Support BI Auditability

Helping EDC users to answer questions like

- „Where does this attribute in my report come from and how is it been calculated/transformed?“ (business users, **financial auditors**)

➔ **Data Lineage** & Impact Analysis, Business Terms

VGH

Prio 2 Usecases EDC@VGH (1)

Support BI Operations

Helping BI Operations to keep „SLA‘s“ by avoiding data production problems due to undetected changes in BI source systems, helping BI and source system developers to identify the BI relevance of objects ➔ Change Tracking & Custom Attributes

Helping EDC users to answer easily questions like

- „What do I have to take into account when changing the dependencies of this BI job in my job scheduler to optimize our loading times?“ (BI Operations / Incident Management)

➔ **Data Lineage & Impact Analysis**

VGH

Prio 2 Usecases EDC@VGH (2)

Support BI Development, BI Management & BI CX

Helping EDC users to answer easily questions like

- „What do I have to take into account when changing this attribute in my ETL chain?“ (BI Development / Change Management)
- „What will happen in the BI frontend when I stop loading this old datamart?“ (DWH modernization)

➔ Data Lineage & **Impact Analysis**

Supporting BI Management in analyzing requirements

➔ **Data Lineage & Impact Analysis**

Helping EDC users to simplify their search for relevant assets by defining more filterable criteria

➔ Analyst‘s Categories as Data Type for EDC Custom Attributes, Bulk Import

VGH

Harvested metadata

Some figures of VGH‘s EDC system (April 2022)

EDC Prod (high)	Assets	Resources	Tables (Views)	(View)Columns
Global	**~ 13 Mio.**	**279 loaded**	**~ 12K (+8.6K)***	**~ 305K (+438K)***
DB2 z/OS		29	~ 1.5K (+0)*	~ 32.9K (+0) *
SQL Server		**216**	~ 10.5K (+8.6K)*	~ 281K (+485K) *
Infa PowerCenter		11	**~ 32.6K map.**	
IBM Cognos		**6**	~649 reports	
Business Glossary		3	~280 BTs	

DWH data in
- 1 DB Instance
- Several Databases (~1 per DWH layer)
- Lots of schemata (~1 per Business topic and layer)
- 1 resource per schema => supports EDC operations and to **structure lineage view**
- **No BI metadata** schemata

Slicing resources as most important conceptional / architectural decision @VGH (⇔ licensing resource based)

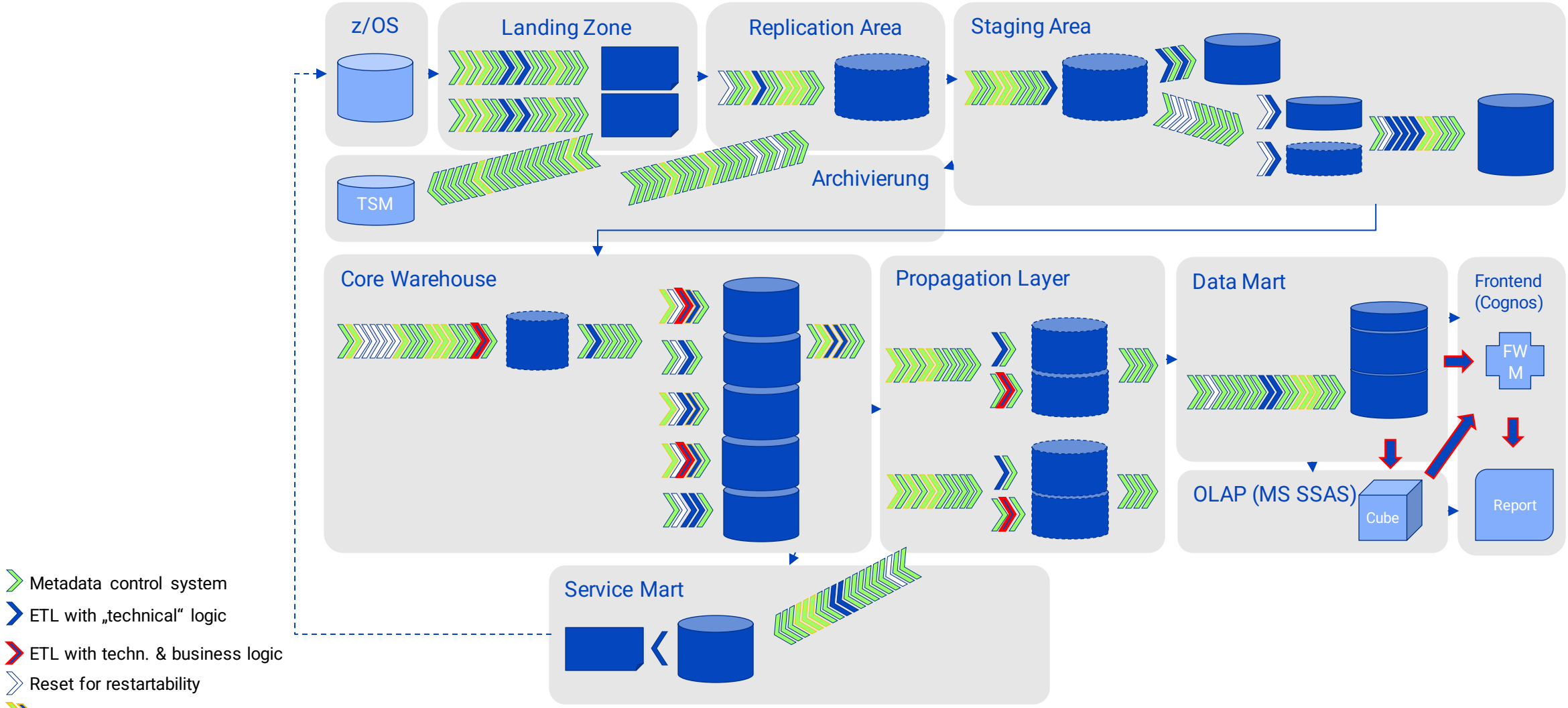
Only a few and use case based ingested to **avoid overwhelming lineage**

Many instances (shortcuts) of reusable objects

* All tables, but no views profiled

VGH

Where does an attribute in a report come from?



Live Demo

Lineage

Thanks for your attention!

Kontakt
Bernhard Link

Schiffgraben 4
30161 Hannover
0511-362-3161
bernhard.link@vgh.de

Questions?

While we answer some of your questions please feel free to also share your thoughts about the session today

Further reading

- Examples of Business versus Physical data structure/lineage representation
  - https://network.informatica.com/docs/DOC-18692

Informatica

Thank You!