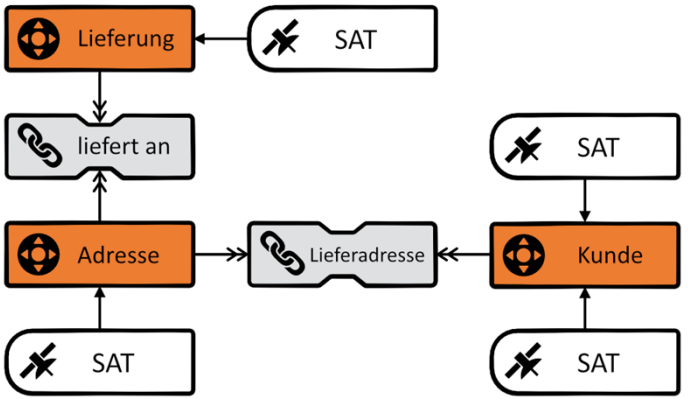
DWH Automation Challenge

- Data Vault Pattern
- Integration
- Faulty source data
- Goals and procedure

DDVUG



Motivation

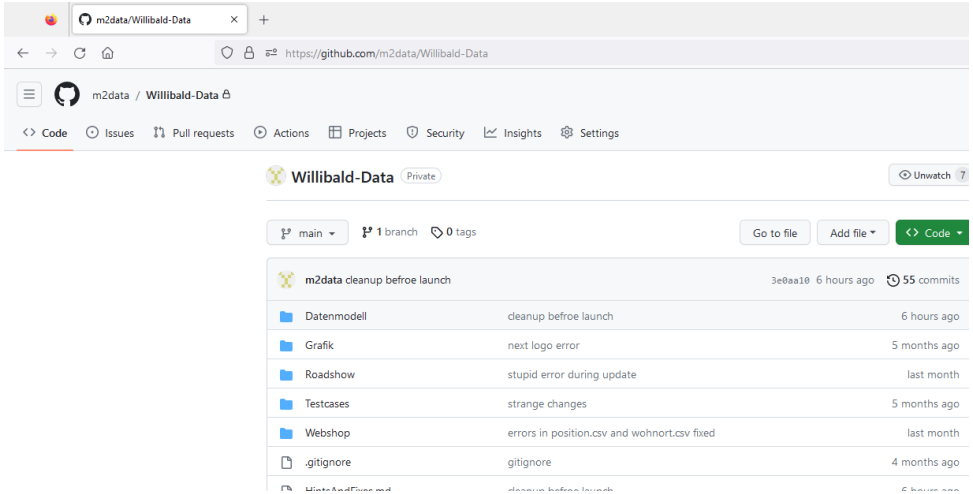
- Today there are many tools for DWH automation and many beautiful statements
- It is difficult for individual DWH teams to understand what works and how it works.
- We want to conduct a test with 4 different DWH automation tools for DWH automation based on Data Vault in order to
  - Draw more attention to the topic of Data Vault and DWH Automation
  - Radically streamline and simplify the tool selection process with materials and criteria catalogues for tool selection
- The test is always positive
  - We do not evaluate, we present by means of examples, the conclusions must be drawn by the potential customer
  - Each publication is agreed in advance

Goals

- Implementation of an openly accessible implementation example with 13 „special cases“ (provided by DDVUG)
- Presentation of the results at the TDWI Conference 2023 (20-22 June).
  - The participants present the implementation of their solution
  - Full day: as a parallel track with the presentations of the 4 participants each approx. 60 to 90 min.
- Provision of a web portal with presentation of the results to simplify the tool selection
- (if necessary, extension with further tools or additional test cases)

Test scenario

- The following describes the test scenario
- The descriptions and the test data are available in a GitHub repository
- The data is available under creative commons CC-4.0-BY

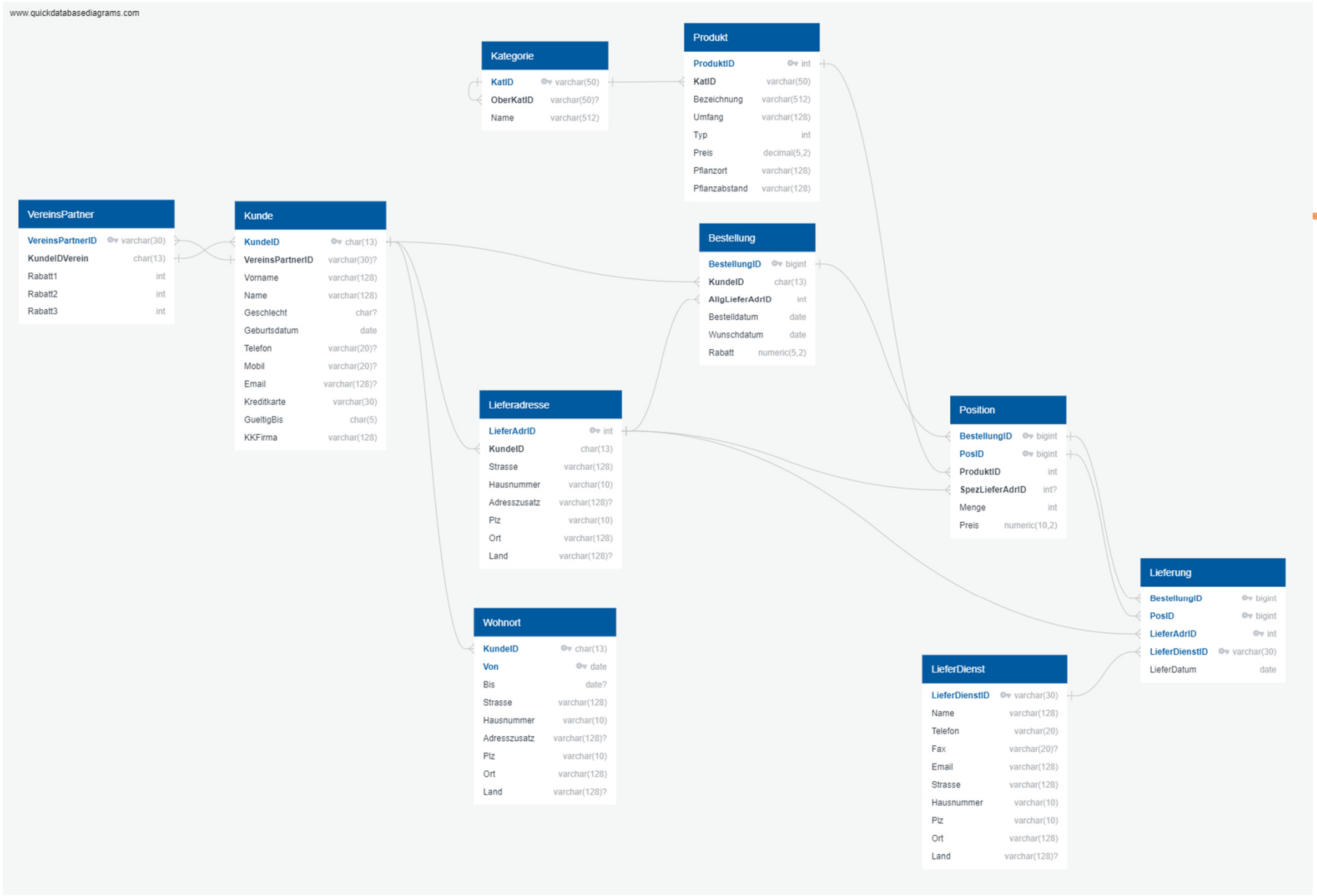


Output data

- The Willibald company is a traditional house and trades in seeds and plants via the internet. In the past, the company sold exclusively through a catalogue. The catalogue was quickly discontinued in 2000, a full 4 years after the webshop was opened. Willibald was the first plant supply shop on the internet and is still proud of it today.

- When ordering via the website, one can enter a desired delivery date and this is currently adhered to 90% of the time. Ordering is a very simple process, the customer chooses his products and orders. They can specify a delivery address for each order item.

- Delivery is then made on the desired date if possible. Since shipping plants also includes shrubs and smaller trees, there are a number of delivery services that Willibald uses for delivery.



DDVUG

Data model Willibald Webshop

Output data II

- The association partners are the backbone of Willibald's marketing. Since its foundation in 1926, the Willibald seed and plant trade has had special conditions and discounts for allotment and horticultural associations. For each association there is a contact person among the customers Each customer can register for his association and thus receive the association benefits. The senior boss is convinced that this concept has brought the Willibald company through all crises.

- Twice a year, the Willibald seed and plant trade goes on a roadshow. During this roadshow, the allotment and horticultural associations are visited with a truck full of seeds and plants. It will organise a small festivity and sell diligently. 2% of the turnover from this truck is donated directly to the association. For Willibald, this is a very good opportunity to get the seasonal goods to the customer before they expire. Since the beginning of the roadshow, no seasonal goods have had to be composted.

- From the roadshow, the data comes from the cash register system. Each turnover is clearly assigned to an association partner. The customer can enter his customer number. Unfortunately, only about 20% of the customers do this. This means that not all of these sales can be assigned to a customer.

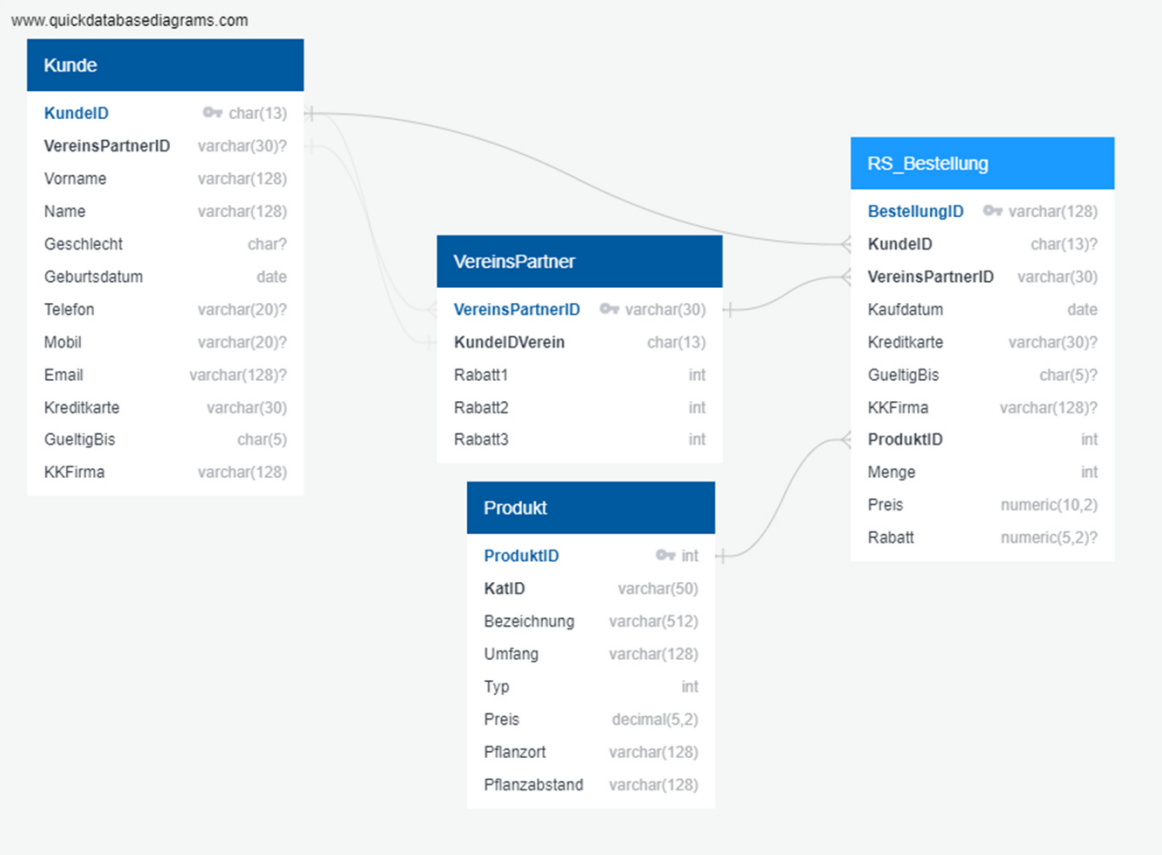
Roadshow table

Only the ORDER table (Bestellung) is delivered. CUSTOMER (Kunde), PRODUCT (Produkt) and ASSOCIATION PARTNER (VereinsPartner) are copies from the webshop.

The order ID (BestellungID) consists of 'RS' and a consecutive number. It is disjunctive to the order ID from the webshop.

The data is delivered in a table. The attributes OrderID (BestellungID), CustomerID (KundeID), AssociationPartnerID (VereinsPartnerID), CreditCard (Kreditkarte), ValidFrom (GueltigBis) and CCCompany (KKFirma) are kept redundant and serve as header information for the rest of the attributes (or actually the items).

The header attributes are always the same for all positions. So far, the data quality is right here.



Cover Data Vault Pattern

The data model contains typical cases / problems / challenges / patterns for the construction of a data vault.

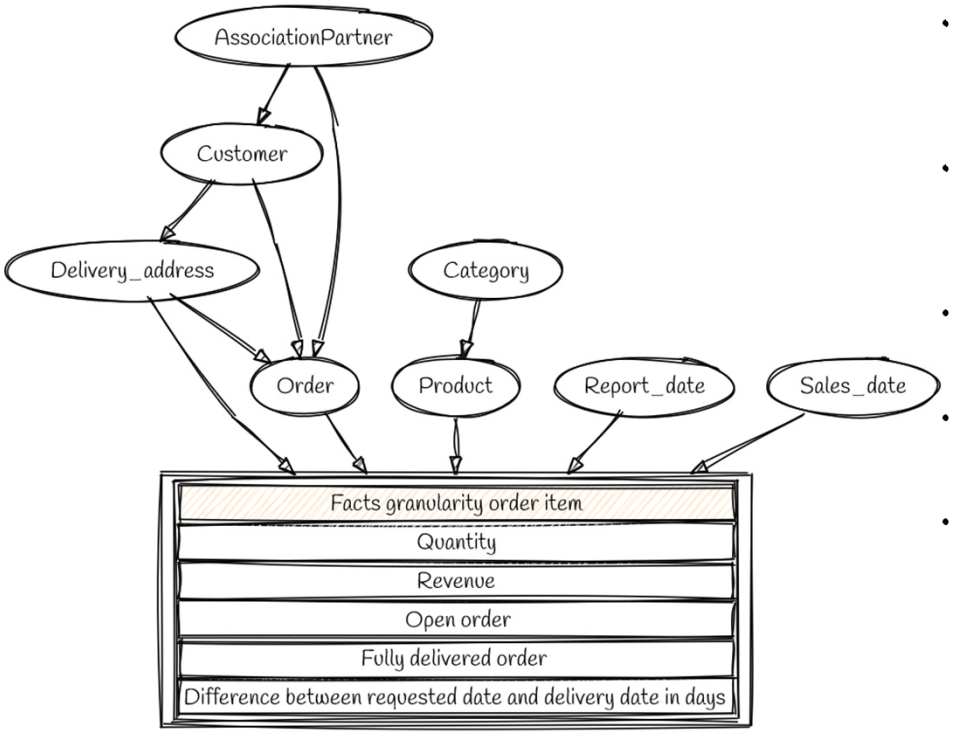
This does not mean that you necessarily have to use the solution pattern suggested here. It is sufficient to solve the underlying situation in such a way that the loading tables can be restored.

A broad coverage of the typical solution patterns is nevertheless desirable.

Hultgren Modelling the agile data warehosue with data vault	Linstedt / Olschimke Building a scalable data warehouse with data vault 2.0	Linstedt (DataModelling Specification 2.0.2) http://datavaultalliance.com/
Hub	**Hub**	**Hub**
Link	**Link**	**Link**
Same As Link (SAL)	Same-As Link (SAL)	Same-As Link (SAL)
Hierarchical Link (HAL)	Hierarchical Link (HAL)	Hierarchical Link (HAL)
	Non-Historized Link	Non-Historized Link
	Dependend Child Link	
		Deep Learning Link
Satellite	**Satellite**	**Satellite**
Multi-Valued Satellite	Multi-Active Satellite	Multi-Active Satellite
(but not recommended)	Status Tracking Satellite	System Driven Satellite
	Effectivity Satellite	Effectivity Satellite
	Record Tracking Satellite	Record Tracking Satellite
	Insert-only Satellite	

Overview of the Data Vault Patterns in the order of their publication.

The evaluation data model



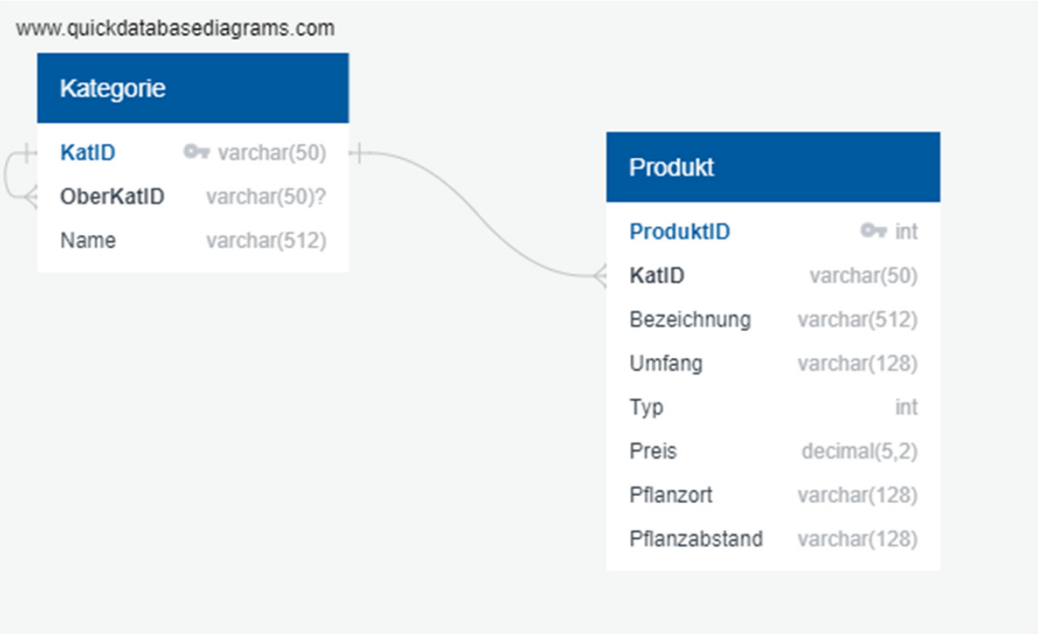
The desired report for Willibald includes the following key figures on granularity of key figures:

- Quantity
  The quantity of products ordered per product. This key figure is taken directly from the order item.
- Revenue
  The amount of money to be paid for the quantity of product ordered. For this purpose, the quantity and price from the order item are multiplied and then the discount is deducted.
- Open order (Bestellung)
  Set 1 to indicate briefly whether this order is still open.
- Completely delivered order (Bestellung)
  Set 1 to indicate briefly whether this order has already been delivered in full.
- Deviation between requested date and delivery date in days.
  If each item has been delivered, the deviation is calculated from the requested date (Bestellung.Wunschdatum) and the last delivery date (Lieferung.LieferDatum). If the delivery was made before the requested date, the deviation is negative. Willibald wants to deliver on time, because too early deliveries also cause problems for the customer (care of seedlings).

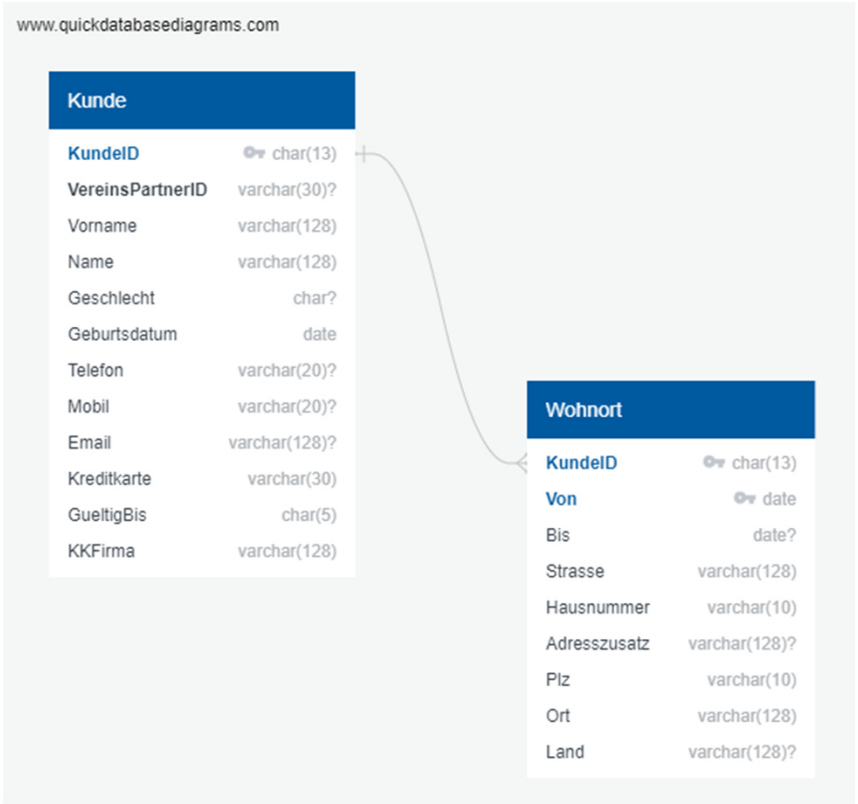
Test criteria

- The Yedi Test is used for successful loading of the Raw Vault:
  - The source tables are restored from the Raw Vault and compared with the originals
  - There must be no deviation
- For the test of the evaluation:
  - the specified values must be supplied in the evaluation

Test cases in the data model hierarchical link



Test cases in the data model multi-active satellite

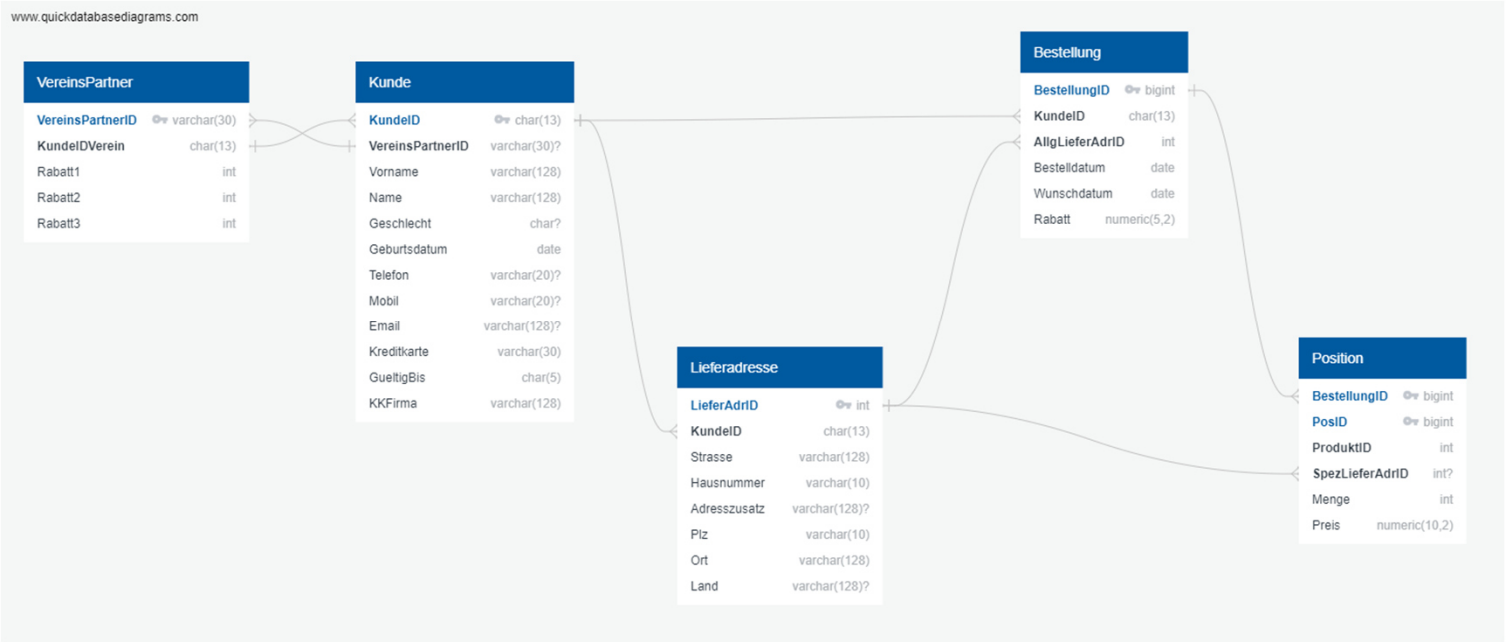


Test cases in the data model
Identifying Relationship and Driving Keys

The relationship between ORDER (Bestellung) and POSITION (Position) cannot change. The key situation makes every change a deletion and a new creation.

All other relationships can change. The test cases are all implemented on the foreign key in CUSTOMER (Kunde) to ASSOCIATION PARTNER (VereinsPartner). The following situations occur here:

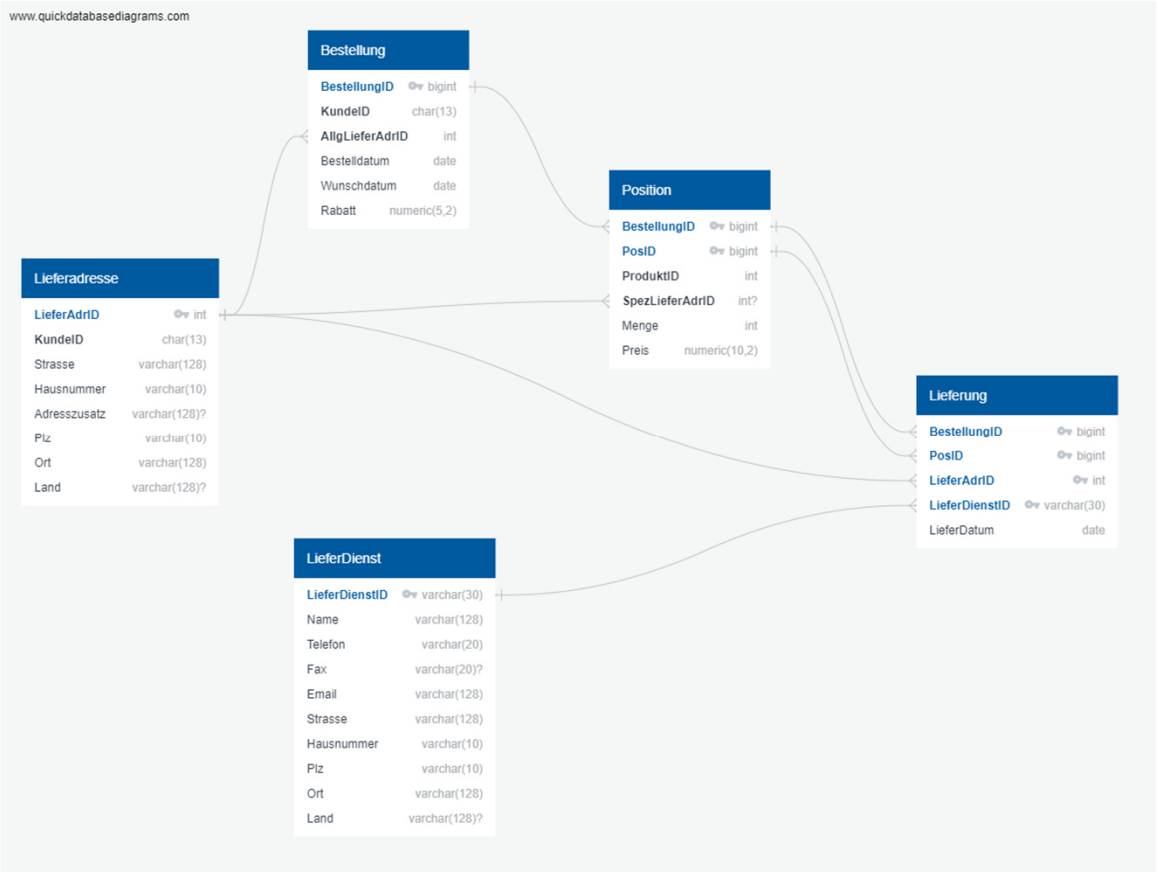
- the foreign key is optional and therefore also NULL
- the foreign key changes between ASSOCIATION PARTNERs
- the foreign key changes from „valid“ to „invalid“ - and in some cases then even back to „valid“ again



Test cases in the data model
m:n table without own key

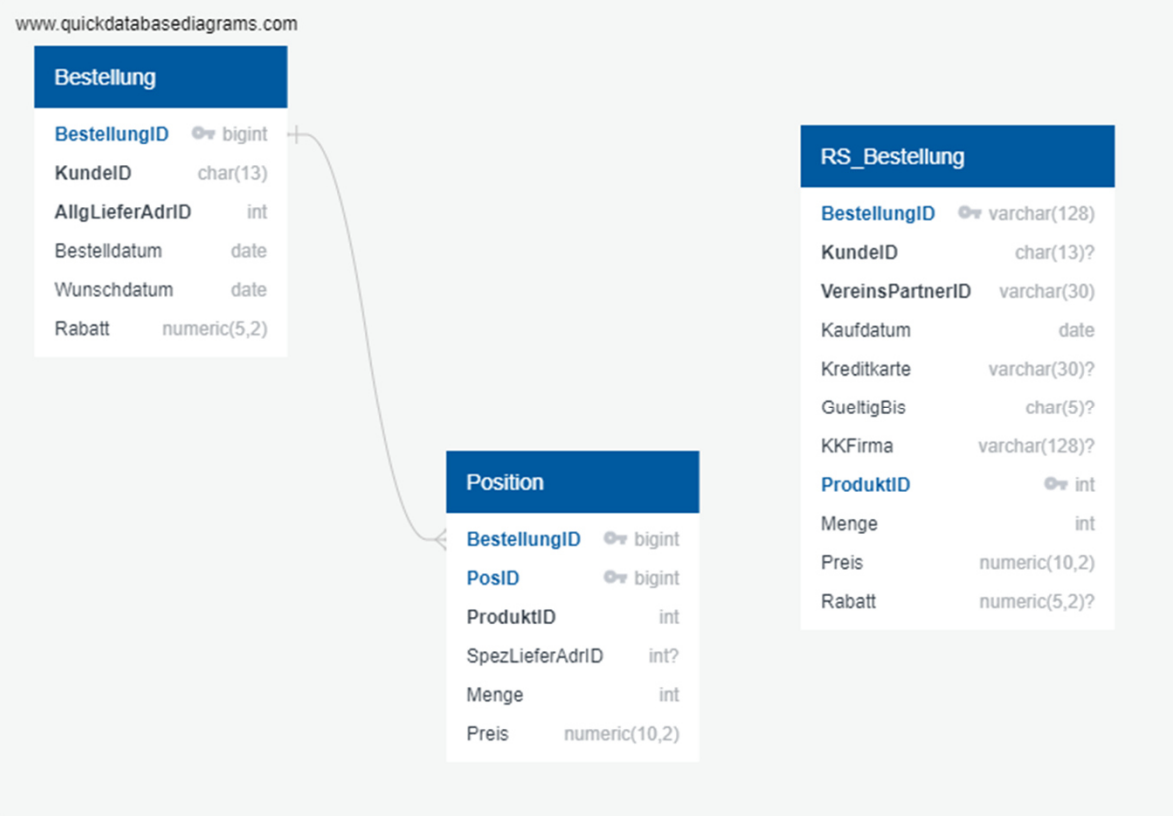
The table of DELIVERIES (Lieferung) does not have its own primary key and usually only occurs once, as only successful deliveries are transmitted to the DWH.

This can be solved in several possible ways: keyed instance, transactional link, dependent child link, ...



Test cases in the data model
Integration of the order

- We have 2 sources. These are to be integrated into ORDER (Bestellung) and POSITION (Position) Hubs.
- The primary key for order (BestellungID) of both systems are simply incremented. The Roadshow has an additional 'RS‘ in for safe differentiation.
- One could argue that integration should take place in the Business Vault. The point here is to present the capability for early integration.



Test cases in the data model (Historicised) reference table

DDVUG

- There is one
  - Reference table
  - historicised reference table

- Both are available as csv file

Test cases in the data

- Duplicates in the loading data
  - There are 2 types of duplicates in PRODUCT (Produkt). In the first case, it is an actual duplicate (ProduktID 20), all attributes are the same. In the second case, the attributes contradict each other (ProduktID 21).
  - Here we just want to see how it is dealt with.

- Rows without business key
  - In DELIVERY SERVICE (LieferDienst), there are records with valid values without a key. Here, too, the only question is how to deal with this.

- Changes in CUSTOMER (Kunde)
  - A very simple test case, the data in the customer (KundeID 107) is changed to a value in delivery 2 and get the values from delivery 1 again in delivery 3.

- Deletions in CUSTOMER (Kunde)
  - There are cases in the customer where the customer (KundeID 70) was deleted in the second delivery and reappears in the third delivery.

Test cases in the data II

- DELIVERY ADDRESSES (Lieferadresse) without CUSTOMER (Kunde)
  - The first delivery contains delivery addresses for which there is no record with the same customer ID (KundeID) in CUSTOMER (Kunde - KundeIDs 999, 998 and 997).

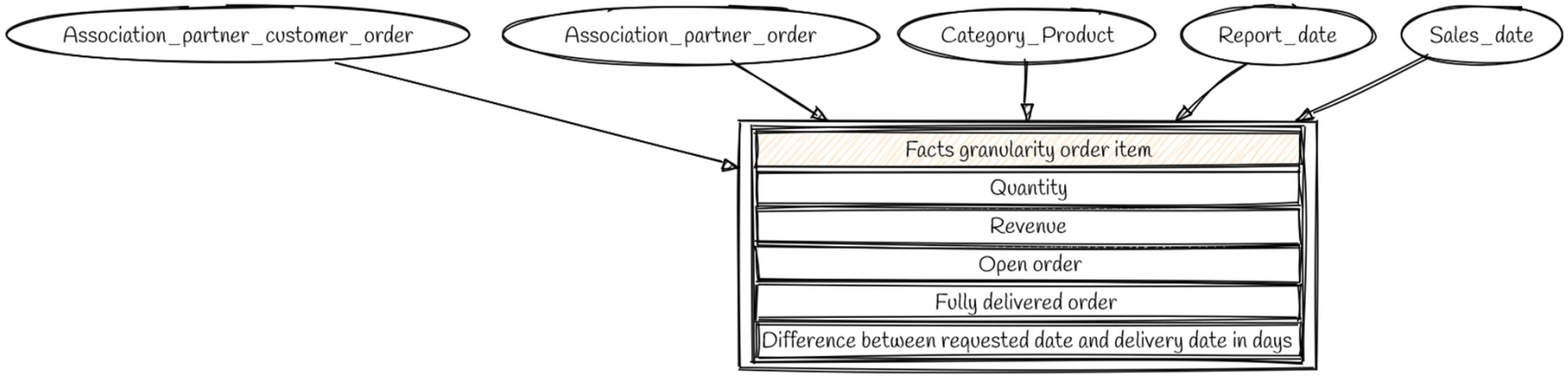
- Deletions of ORDERS (Bestellung)
  - The orders are relevant for counting and are deleted during the dates of deliveries.
  - Between period 1 and 2 the orderIDs (BestellungID) 99, 220 and 465.
  - Between period 2 and 3 the orderIDs (BestellungID) 1470 and 1288.

- Changes in the dimensions
  - The hierarchy of the product CATEGORY (Kategorie) changes completely with both deliveries. So we have 3 different product hierarchies. These are to be displayed as as-what at the respective reporting time.

Key figures as result control

The key figures are controlled:

- Quantity
  The quantity of products ordered per product. This key figure is taken directly from the order item.
- Revenue
  The amount of money to be paid for the quantity of product ordered. For this purpose, the quantity and price from the order item are multiplied and then the discount is deducted.
- Open order
  Set 1 to indicate briefly whether this order is still open.
- Fully delivered order
  Set 1 to indicate briefly whether this order has already been delivered in full.
- Difference between requested date and delivery date in days.
  If each item has been delivered, the deviation is calculated from the requested date (Bestellung.Wunschdatum) and the last delivery date (Lieferung.LieferDatum). If the delivery was made before the requested date, the deviation is negative. Willibald wants to deliver on time, because early deliveries also cause problems for the customer (care of seedlings).



Business Rules

1. Standardise ORDERS (Bestellung) of the ASSOCIATION PARTNERS (VereinsPartner)
   The ROADSHOW ORDERS (RS_Bestellung) are directly linked to the ASSOCIATION PARTNERS (VereinsPartner). This must now be done for the orders of the association partner from the webshop. To do this, all orders of this customer are linked directly to the association partner using the association partner ID (VereinsPartnerID) of this CUSTOMER (Kunde).

2. Roadshow: assigning ORDERS (RS_Bestellung) to CUSTOMERS (Kunde)
   If the customer ID (KundeID) is missing in the ROADSHOW ORDERS (RS_Bestellung), the credit card, CC-Company (KKFirma) and the valid-to (GueltigBis) can be used to identify the customer.

Overarching functions

- In addition to data preparation, further functionalities are needed. Ideally, these should be covered by the automation tool or - if not - at least an interface to such a tool should be available.
- The following functionality can be represented by one or more tools. For the time being, this is only about the representation of the functionality.

Overarching functions Block 1

DDVUG

- Data Lineage
  A representation of the dependencies in data preparation. Either at attribute level or at entity level. This makes it possible to understand how changes to the interface affect the key figures or from which data sources a key figure is composed.
  The data for the lineage should not have to be documented separately, but should essentially result from the implementation and be maintainable with little effort in the course of further development.

- Orchestration
  The individual processing steps and their interdependencies are defined in the orchestration.

- Scheduling
  The control of the preparation processes with the possibility of parallelising the loading and executing it in a time-controlled manner. Ideally, load balancing can be carried out.

Overarching functions Block 2

- Error Handling
  Handling of errors and notification that an error has occurred. The Data Vault also loads erroneous data, so errors should be reported in an 'error mart' or similar.
- Deployment
  Additional Loads
  Destructive changes - rebuilding data structures
  Partitial changes

Additional info

- Supported databases
- Installation Tool onPrem/Saas/both

Translations of Entities

	Translation	Description
• Bestellung	Order	List of all Orders
• Kategorie	Category	Plant Categories – including several aggregation levels
• Kunde	Customer	List of all Customers
• Lieferadresse	DeliveryAddress	All delivery addresses per customer
• LieferDienst	DeliveryService	List of Services dealing with delivieries (e.g. DHL)
• Position	Position	Order Positions
• Produkt	Product	List of all Products
• VereinsPartner	AssociationPartner	List of Gardening Clubs – they are also Customers
• Wohnort	LivingAddress	History of addresses per Customer