Projektarbeit
Real Estate
Schmitt AG

Gruppe2

Jann Subbras, Ihsan Bilici, Tat Dan Tran, Christos Terzoglou

Inhalt

1 Scenario

Scenario Schmitt AG

Branche: Immobilien

Bereich: Vermittlung von Immobilien

Die Schmitt AG möchte ihr Geschäftsfeld erweitern und in neue Geschäftsbereiche investieren. Deswegen hat die Schmitt AG vor 3 Monaten auch einen Real Estate Bereich gegründet, deren Hauptaufgabe es ist, Immobilienobjekte zu vermitteln. Für den Betrieb der neuen Abteilung wird ein neues IT-System benötigt, zu dem eine Relationale Datenbank für die Datenverwaltung gehört. Mit den Daten sollen abgesehen von den Standardabfragen, auch analytische Auswertungen möglich sein. Aus dem Grund wird auch ein Data Warehouse angebunden, mit dem solche Analytische Abfragen realisiert werden.

1 Scenario

Entitäten

Entity Relationship Modell (OLTP) für die operative ERP Datenbank
Stammdaten/Bewegungsdaten

- Abbildung der ChanelArt
- Abbildung der Datum
- Abbildung der Expose
- Abbildung der ExposeVersand
- Abbildung der Kauf
- Abbildung der Kontakt
- Abbildung vom KontaktArt
- Abbildung der KundenArt
- Abbildung vom Lage
- Abbildung der Lead
- Abbildung vom Objekt
- Abbildung der ObjektArt
- Abbildung der Person
- Abbildung der PersonenArt
- Abbildung der PersonKauf
- Abbildung der PersonKontakt
- Abbildung der PersonObjekt
- Abbildung der PersonRolle
- Abbildung der PersonTermin
- Abbildung der Termin
- Abbildung der TerminArt
- Abbildung der Wunsch

Data Vault Modell ist ähnlich strukturiert wie das OLTP –Modell, daher verwenden wir im ersten Projektschritt die selben Entitäten
Dimensionen / Fakten

- Abbildung der ChanelArt
- Abbildung der Datum
- Abbildung der Expose
- Abbildung der ExposeVersand
- Abbildung der Kauf
- Abbildung der Kontakt
- Abbildung vom KontaktArt
- Abbildung der KundenArt
- Abbildung vom Lage
- Abbildung der Lead
- Abbildung vom Objekt
- Abbildung der ObjektArt
- Abbildung der Person
- Abbildung der PersonenArt
- Abbildung der PersonKauf
- Abbildung der PersonKontakt
- Abbildung der PersonObjekt
- Abbildung der PersonRolle
- Abbildung der PersonTermin
- Abbildung der Termin
- Abbildung der TerminArt
- Abbildung der Wunsch

1 Scenario

ETL

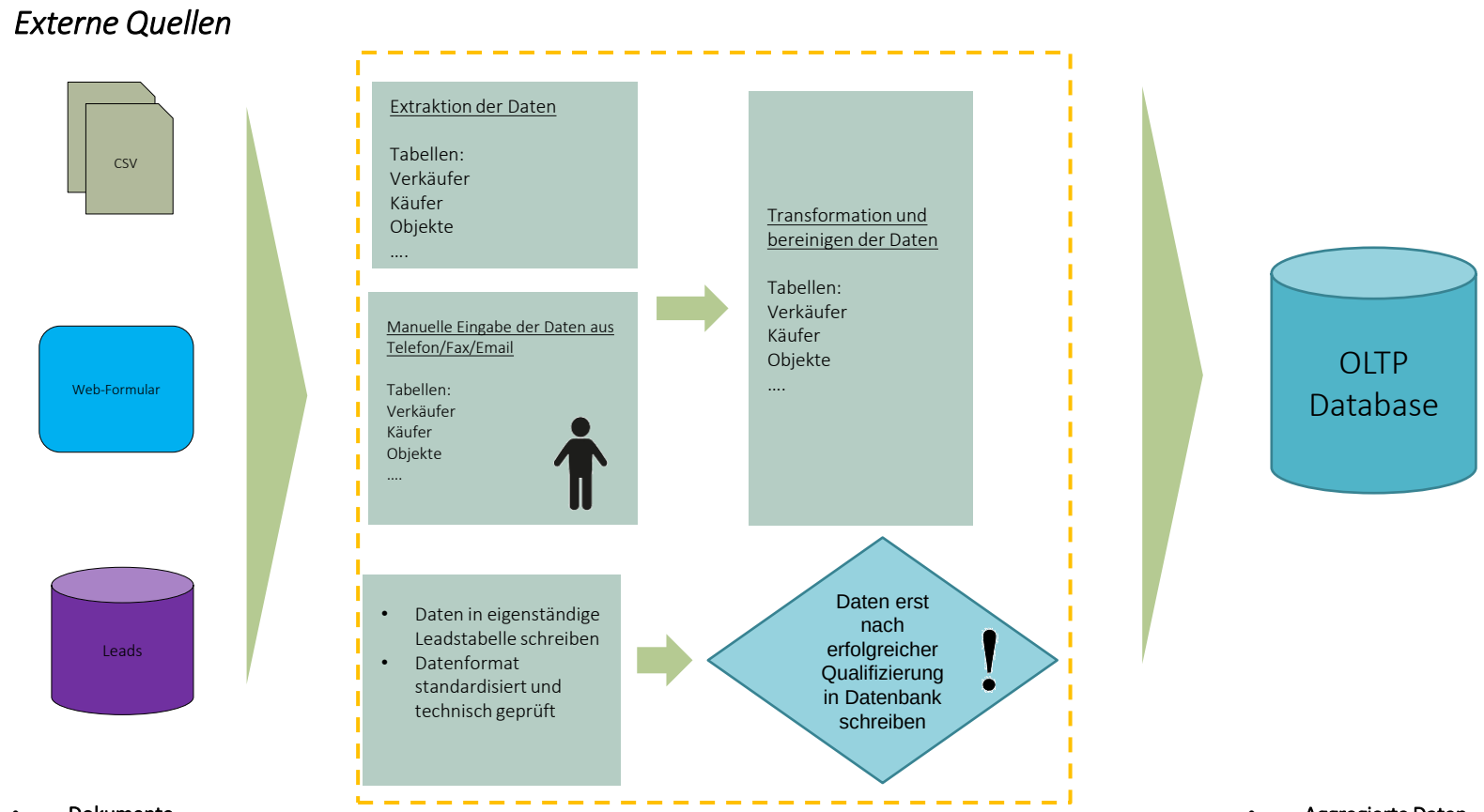
- Erstellung ETL Prozess für die Personendaten von der Operativen Datenbank ins Data Vault
- Erstellung ETL Prozess für die Kundendaten von der Operativen Datenbank ins Data Vault
- Erstellung ETL Prozess für die Objektdaten von der Operativen Datenbank ins Data Vault
- Erstellung ETL Prozess für die Lagedaten von der Operativen Datenbank ins Data Vault
- Erstellung ETL Prozess für die Wunschdaten von der Operativen Datenbank ins Data Vault
- Erstellung ETL Prozess für die Vertragsdaten von der Operativen Datenbank ins Data Vault
- Erstellung ETL Prozess für die Lagedaten von der Operativen Datenbank ins Data Vault
- Erstellung ETL Prozess für die Termindaten von der Operativen Datenbank ins Data Vault

Datenquellen

- Alt-Daten
- CSV-Dateien
- Webformular
- Leads
- Telefon/Fax/E-Mail (Erfordern manuelle Eingabe)

2 Architektur OLTP

OLTP Architektur



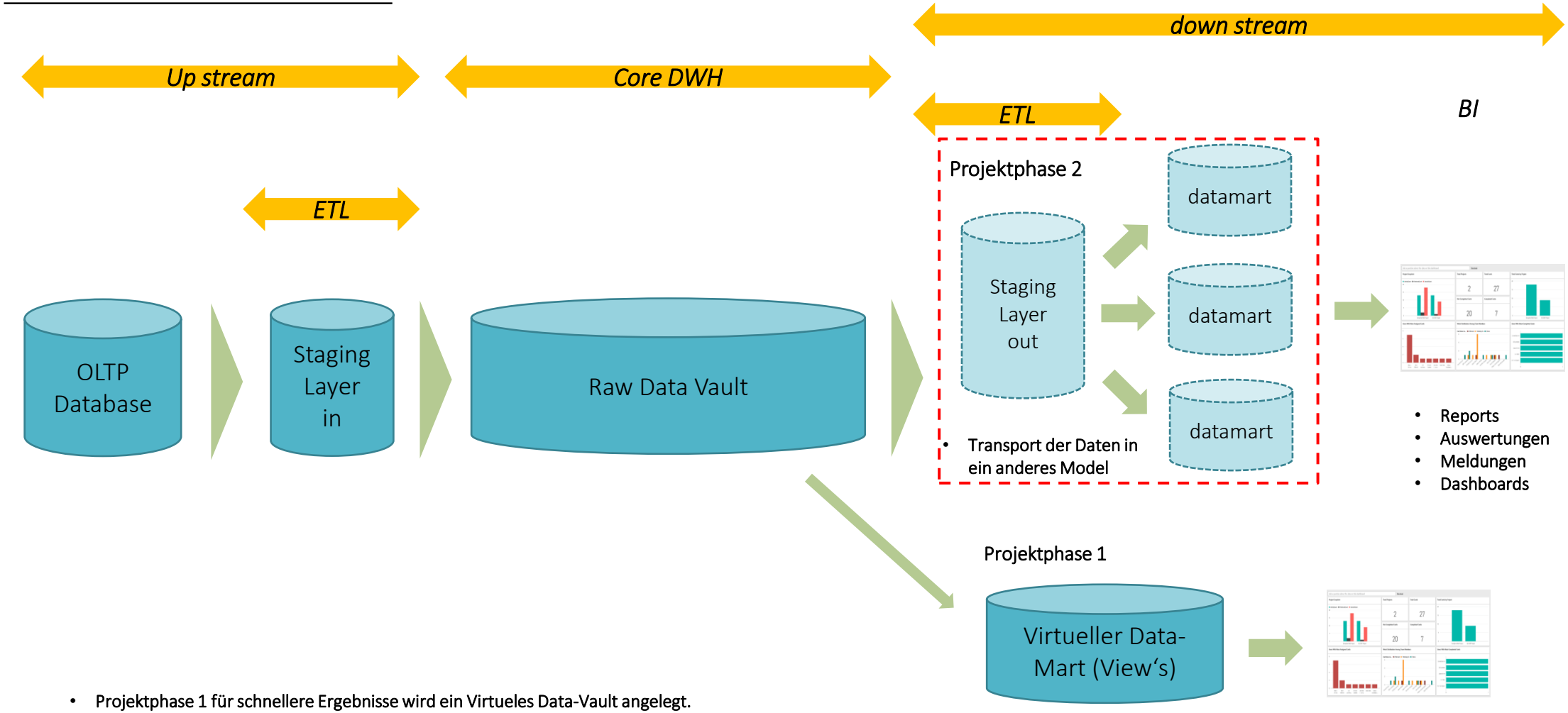
- Dokumente
- Operative Datenquellen

- Zusammenführen Daten aus verschiedenen Quellen
- Datenstruktur wird festgelegt
- Erzeugte Tabellen werden in Operatives Data Storage geschrieben

- Aggregierte Daten

3 Architektur Data Warehouse

Data Warehouse Architektur



- Projektphase 1 für schnellere Ergebnisse wird ein Virtueles Data-Vault angelegt.
- Projektphase 2 wird auf eine Klassische Datamart Struktur angepasst werden

4 Density und Datenvolumen

Für die Berechnung des Speicherbedarfs wird eine Prognose für Zeitraum von 1, 5 und 10 Jahre getroffen.
Berücksichtigt werden aus der Datenbank die Tabellen mit den meisten Merkmalen.

- Tabellen mit größten Speicherbedarf sind : Lage, Objekte, Person, Wunsch
- Zu den wichtigsten Nebentabellen gehören Datum, VersandExpose, Kauf, Kontakt

Quantität:

- 1000 Anfragen pro Tag
- Abteilung läuft bereit seit 3 Monaten : 3 * 30 Tage = 90 Tage

Für eine Berechnung des benötigten Speicherbedarfs werden folgende annahmen getroffen:

- 4 Byte für jeden Primary- und Foreign-key und Integer
- 50 Byte für Text
- 100 Byte für einen Link

4 Density und Datenvolumen

<u>Haupttabellen:</u>

Tabelle Person:
2PK +2FK+16 Merkmale = 2* 4 + 1 * 4 + 15 * 50 = 762 * 1000 = 720000 Bytes/Tag

Für die ersten 3 Monaten haben sich 90 * 720000 = 64800000 /1000000 = 64,8MB an Personendaten angesammelt.

Tabelle Lage:
2PK + 6 Merkmale = 1* 4 + 6 * 50 = 304 * 1000 = 304000 Bytes/Tag

Für die ersten 3 Monaten haben sich 90 * 304000 = 27360000 /1000000 = 27,36MB an Lagedaten angesammelt.

Tabelle Objekte:
1PK +3FK+11 Merkmale = 1* 4 + 3 * 4 + 8 * 50 + 3 * 4 = 428 * 1000 = 428000 Bytes/Tag

Für die ersten 3 Monaten haben sich 90 * 428000 = 38520000 /1000000 = 38,52MB an Objektdaten angesammelt.

Tabelle Wunsch:
1PK + 3FK + 6 Merkmale = 1* 4 + 3 * 4 + 2 * 50 + 2 * 4 = 124 * 1000 = 124000 Bytes/Tag

Für die ersten 3 Monaten haben sich 90 * 124000 = 11160000 /1000000 = 11,16MB an Wunschdaten angesammelt.

4 Density und Datenvolumen

Nebentabellen:

Tabelle Datum:
1PK + 4 Merkmale = 1* 4 + 3 * 4 + 1 * 50 = 66 * 1000 =66000 Bytes/Tag

Für die ersten 3 Monaten haben sich 90 * 66000 = 5940000 /1000000 = 5,94 MB an Datumsdaten angesammelt.

Tabelle ExposeVersand: für 1000, 5000, 10000
1PK + 3PK + 1 Merkmal = 1* 4 + 3 * 4 + 1 * 100 = 116 * 1000 = 116000 Bytes

Beim Versand nicht das Expose selbst sondern ein Link in der Tabelle gespeichert 116000 /1000 = 116Kb Versanddaten angesammelt.

Tabelle Kauf:
1PK + 2FK + 1 Merkmal = 1* 4 + 2 * 4 + 1 * 50 = 62 * 1000 = 62000 Bytes/Tag

Für die ersten 3 Monaten haben sich 90 * 62000 = 5580000 /1000000 = 5,58 MB an Kaufdaten angesammelt.

Tabelle Kontakt:
1PK + 3FK + 1 Merkmal = 1* 4 + 3 * 4 + 1 * 50 + 2 * 4 = 124 * 1000 = 66000 Bytes/Tag

Für die ersten 3 Monaten haben sich 90 * 66000 = 5940000 /1000000 = 5,94 MB an Kontaktdaten angesammelt.

4 Density und Datenvolumen

- Für 1 Exposé (PDF mit Bildern und Text) wird ein Speicher von jeweils 500kb benötigt.
- Für jedes Objekt existiert ein Exposé.
- Für die Menge von Exposés wird die Annahme getrogen das es >1000 in den ersten Jahren sein werden.
- Für die Berechnung wird daher eine Anzahl von 1000, 5000, 10000 für Zeiträume 1-10 Jahren verwendet.
- In den Tabelle „ExposeVersand“ wird der Link zu dem jeweilige Exposé gespeichert.

- Benötigter Speicher für ein Portfolio von 1000 Exposés** im ersten Jahr: 500kb = 500000 * 1000 Exposés = 500MB

Haupttabellen

Tabellen/Zeit	1 Jahr	5 Jahre	10 Jahre
Lage	27,36 MB	27,36 MB	27,36 MB
Objekte	38,52 MB	192,6 MB	385,2 MB
Person	259,2 MB	1,296 GB	2.592 GB
Wunsch	11,16 MB	55,8 MB	111,6 MB
Summe	**350 MB***	**1,6 GB***	**3,2 GB***

Nebentabellen

Tabellen/Zeit	1 Jahr	5 Jahre	10 Jahre
Datum	5,94 MB	29,7 MB	59,4 MB
ExposeVersand	116 KB	580KB	1,16 MB
*Expose PDF***	*500MB*	*2,5GB*	*5GB*
Kauf	5,58 MB	27,9 MB	55,8 MB
Kontakt	5,94 MB	29,7 MB	59,4 MB
Summe	**518 MB***	**2,6 GB***	**5,2 GB***

Den Notwendigen Speicher stellt die IT-Abteilung der Schmitt AG bereit
und wird von der Abteilung Real Estate angemietet!

* Die Summen gerundet

5 Data Quality

Aktuelle Probleme, die in der vom Kunden bereitgestellten Tabellen festgestellt wurden

- **Anrede**: Herr, Frau oder Firma. Aber Familie befindet sich unter dieser Kategorie.
- Fehlende Primärschlüssel (nicht vorhanden)
- **Vorname**: Schreibfehler (Vorname = P. , Nachname =Aydemir)
- **Straße**: Schreibfehler (Straße = B ,)
- **Ansprechpartner**: Für Firma gibt es fehlende Ansprechpartner (Firma Gauch)
- **Tel FN :** Alle Telefonnummern sind 9 Ziffern außer einer.
- **Telefon privat** (Tabelle Käufer) : 12 Ziffern
- **Preisvorstellung von bis (Preisminimum / Preismaximum)** : Für diese beiden unterschiedlichen Werte werden eine einzige Spalte erstellt (150000-180000) und manchmal nur ein Wert (285000)
- **E-Mail** : Fehlende Daten – alle sind leer
- **gewünschte Kontaktart** (telefonisch, per eMail, Whatsapp, etc.) Aber wurde ‘Telefon‘ geschrieben.
- **Falsche Eintrag für Datum** (Tabelle Käufer) **:** ‘3.11.2021<u>2</u>11‘
- **Art des Objektes** (1FH, Mehrfamilienhaus, etc.) : Einfamilienhaus statt 1FH
- **Falsche Eintrag:** Expose zugesendet vor dem ersten persönlichen Telefonat

5 Data Quality

Mögliche Probleme

Scope/Problem	Dirty Data	Reasons/Remarks
Abkürzung, Kryptische Werte	Title = Dr, dr, Doktor	verschiedene schreibweise/Einträge
Unzulässiger Wert	Anrede = ‘Familie‘	Werte außerhalb des Admin Bereichs
Unterschiedliche Repräsentationen	Anrede = Herr, Frau, Familie Anrede = 1,2,3	verschiedene Wertebereiche
Unterschiedliche Repräsentationen	Preis = in Euro Price = in Tausend Euro	verschiedene Einheiten
Eindeutigkeit verletz	PLZ Frankfurt 61200 PLZ München 61200	nicht eindeutige PLZ
Falsche Zuordnung	Ort =Deutschland	Werte außerhalb des Admin Bereichs
Schreibfehler	Ort = Frankfurt Ort = Frankfut	Schreibfehler
Fehlende Werte	Telefonnummer = 49622112445517 (14-stellig)	Schreibfehler
Fehlende Daten	Ansprechpartner für Firmen	unpflichtfelder
unzuverlässiger Wert	Datum: 32.12.2022	Werte außerhalb des Admin Bereichs
Attribute Abhängigkeit verletzt	Termindatum > Kaufdatum	Abhängigkeiten nicht definieren
Falsche Format	Datum = 18/11/2022 Datum = 18112022	verschiedenes Format

Scope/Problem	Dirty Data	Reasons/Remarks
Unzulässiger Wert	Kauftermin = 18.02.2322	Werte außerhalb des Admin Bereichs
Formatfehler	Kaufpreis = 15.2a5.000 Preis= Float, Preis= String	Formatfehler
Unplausible Daten	Anzahl Zimmer = 150 qm=85	Ausreißer / Schreibfehler
Fehlende Daten	E-Mail, Preisminimum, Preismaximum	unpflichtfelder
Attribute Abhängigkeit verletzt	Preisminimum > Preismaximum	Abhängigkeiten nicht definieren
Schreibfehler	Kontaktwunsch (wann) = (Ab 18) Kontaktwunsch (wann) = (ab 19 Uhr)	Verschiedene schreibweise
Schreibfehler	gewünschte Kontaktart = telefon, telefonisch	Verschiedene schreibweise
Schreibfehler	Art des Objektes = 1FH Art des Objektes = Einfamilienhaus	Verschiedene schreibweise
referentielle Integrität verletzt.	Kauftermin =21.01.2022 ObjektID = 4567	Referenzierte ObjektID nicht vorhanden/schon verkauft
Fehlende Daten	e-mail = null	unpflichtfelder
Schreibfehler	e-mail = bilici@gmail	Verschiedene schreibweise

5 Data Quality

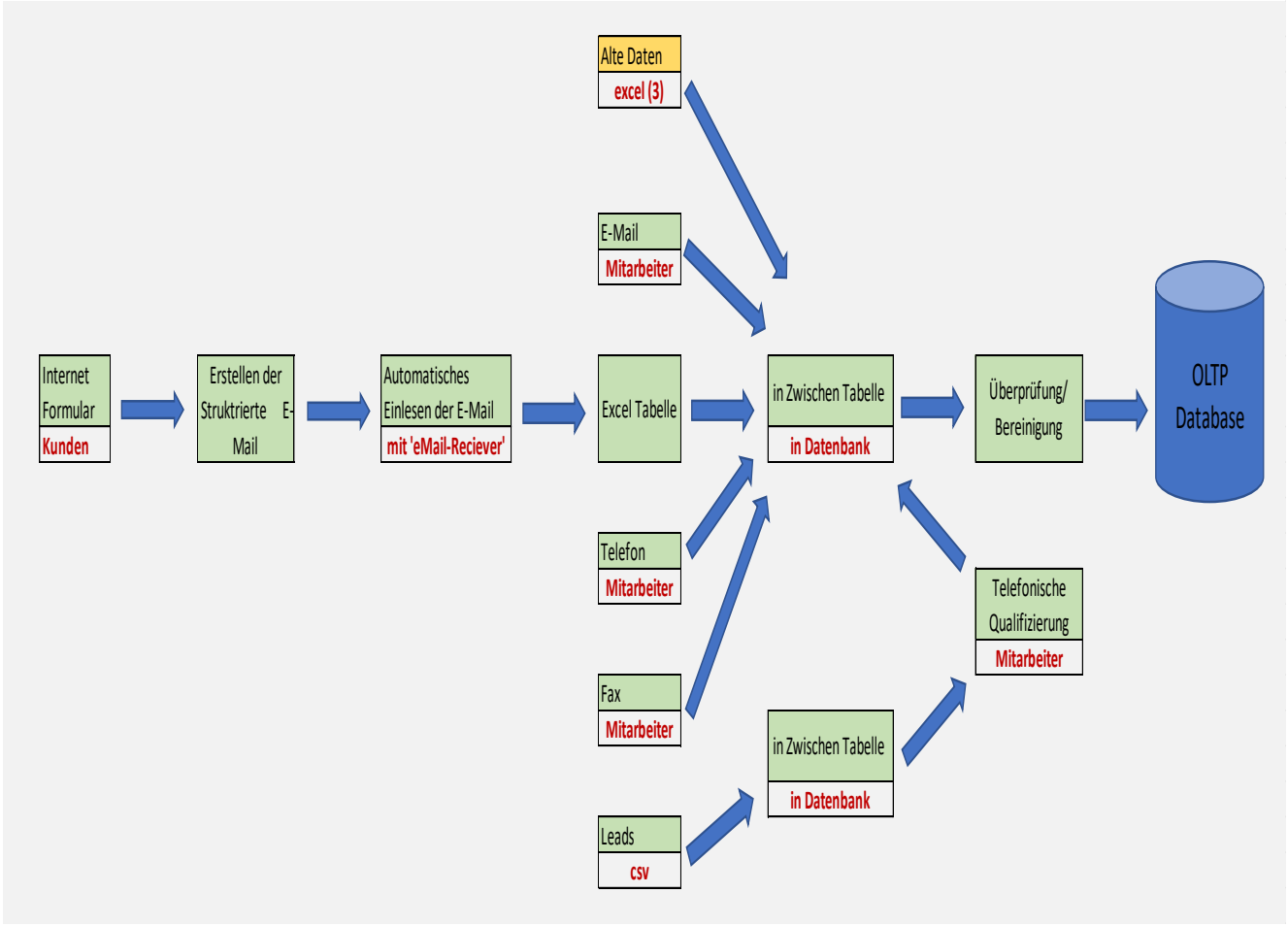
Lösungen zur Vermeidung von den möglichen Problemen

Fehlerart	Lösung	Attribute
Fehlende Daten	Pflichtfelder definieren	E-Mail, Ansprechpartner (für Firmen), Preisminimum, Preismaximum, Vorname, Nachname, Anrede, Anzahl Zimmer, Größe, Kundentypen, Ort, PLZ, Straße, Land, Bundesland, Telefonnummer
Abkürzung	Drop Down Liste	Titel
unzuverlässiger Wert	Drop Down Liste	Anrede, Quelle
unzuverlässiger Wert	Drop Down Kalender Liste	Datum, Kauftermin
Unplausible Daten	Max und min definieren	Anzahl Zimmer, Preisminimum, Preismaximum, Kaufpreis, Provision, Preis, Größe, Zimmer, Hausnummer, ProvisionProzent, verkauft
Schreibfehler	Drop Down Liste	Kontaktwunsch (wann), gewünschte Kontaktart,, Art des Objektes, Kundentypen, Ort

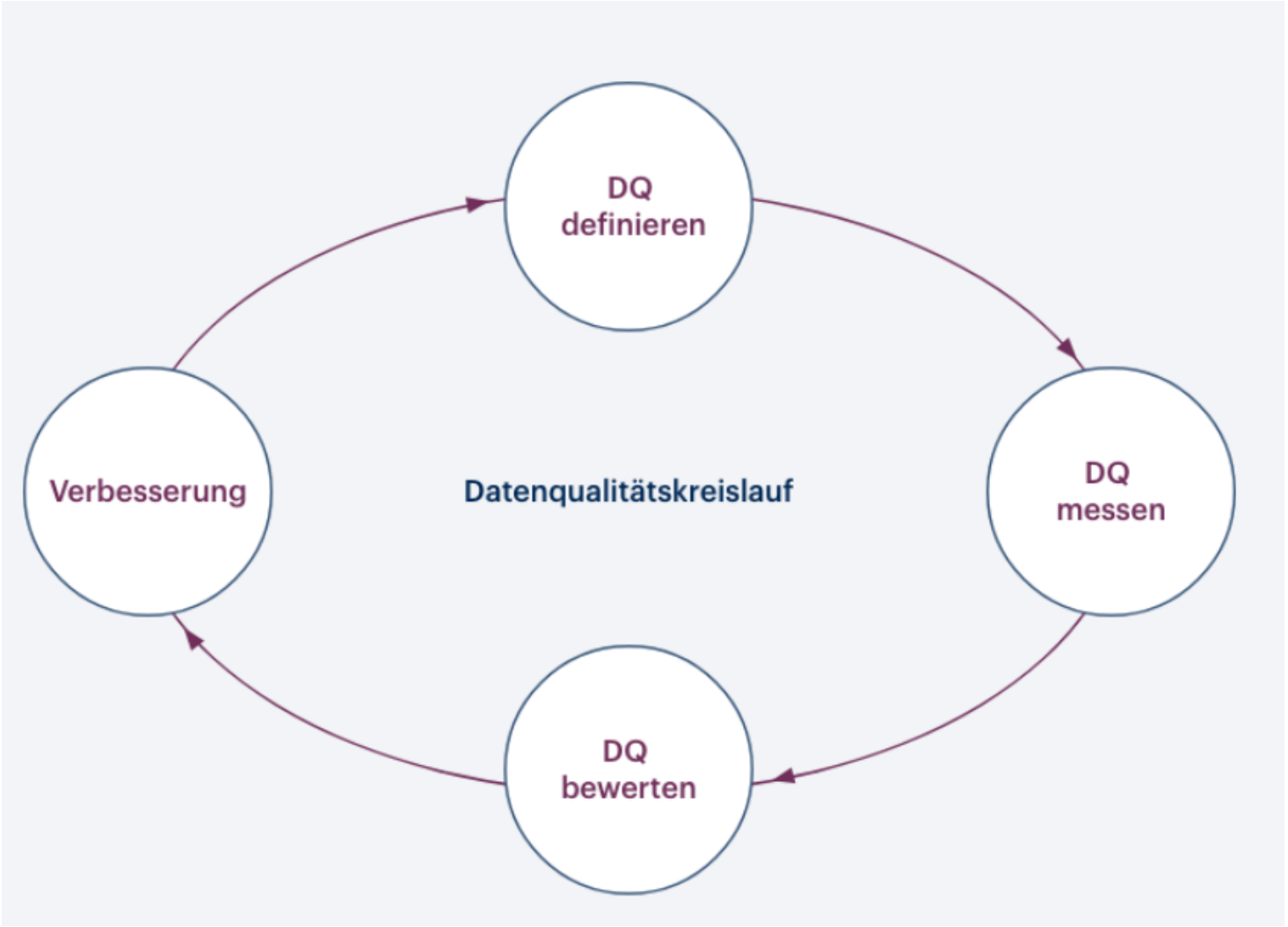
Fehlerart	Lösung	Attribute
Schreibfehler	Einschränkungen definieren	E-Mail, Telefonnummer, Fax, TelefonGeschäftlich, TelefonPrivat, TelefonMobil, ExposeLink
Schreibfehler	einheitliche Groß/Kleinschreibungsform	Vorname, Nachname, Firma
Falsche Zuordnung	Vergleichen mit referenz Tabelle	Ort, PLZ, Lage
Eindeutigkeit verletzt	Lookup/Referenz Tabelle	Ort, PLZ, Straße, Land, Bundesland,
Attribute Abhängigkeit verletzt	Einschränkungen definieren, Identifizieren von ungültigen Werten	Preisminimum, Preismaximum, E-Mail, DatumID(Kauf), DatumID(Termin)
Falsche Format	einheitliches Format für Datums-/Zeit-Angaben	Datum, Jahr, Monat, Tag

5 Datenflüsse

Probleme	Beispiel	Mögliche Lösung
Einlesen	Csv, mit Java App(E-Mail)	Prüfsumme, Anzahl der verarbeiteten Zeilen
unlesbar	Fax	Nachfragen /Fax vermeiden
Manuelle Eingabe (Web Formular /Operativen Systemen)	Tippfehler	Drop Down Liste, interne Prüfprozedur, Pflichtfelder
Fehler bei elektronischer Datenübertragung	File Transfer	Prüfsumme
Automatische Update, bestimmte Programme funktionieren nicht mehr	Mit Java App, Microsoft DDL	Regelmäßig Fehler Protokolle auswerten
Excel Probleme	Leerzeichen, leere Spalten, Tippfehler, inkonsistente Strukturen zwischen gleichartigen Excel Files	Strukturprüfrutine
Beliebige Datenquellen: Fehlende Felder	Fehlender Ort	Lookup in Postdatei, Ort ergänzen



5 Data Quality



6 Dokumentation

- Für die Dokumentation der OLTP Datenbank existiert ein separates Dokument.
- Für die Dokumentation der Datenbank gibt es ein separates Dokument.

7 Datenbanken
7.1 Einführung

Im **Proof of Concept** wird die generelle Machbarkeit des angedachten Projektes überprüft, einige technische Möglichkeiten ausgelotet und an der Realität verprobt.

In dieser ersten Phase eines möglichen Projektes geht es **nicht** um die **exakte Vorwegnahme** künftiger Strukturen und Abläufe.

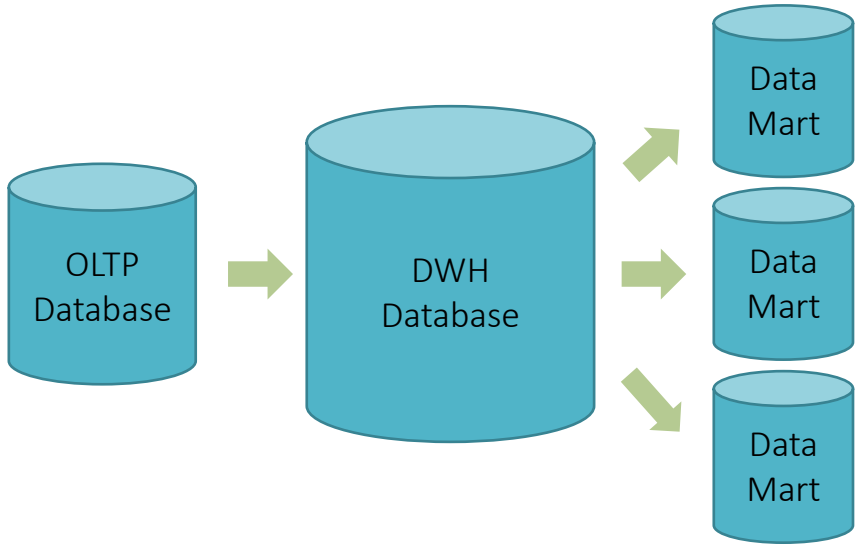
Im Verlaufe eines Projektes werden sich aufgrund inhaltlicher, technischer und organisatorischer Gegebenheiten einiges verändern.

Diesen Veränderungen wird mit einer agilen Vorgehensweise Rechnung getragen.

7.2 Datenbanken

Dieser Teil der Gesamtpräsentation beschäftigt sich ausschließlich mit den Strukturen der zu entwickelnden Datenbanken.

Prozesse, Abläufe, Datenflüsse, ETL-Pipelines u.ä. werden an anderen Stellen behandelt.



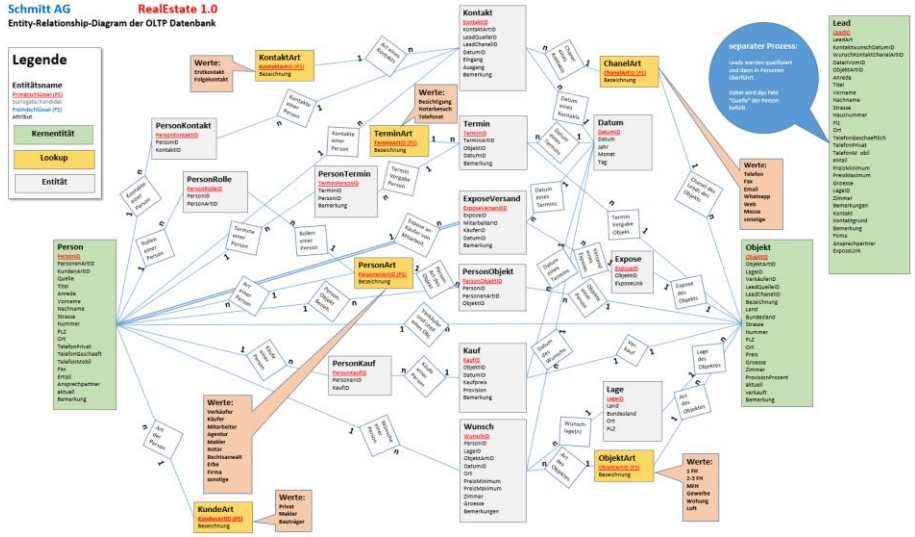
7.3 OLTP

Eine OnLine Transactional Processing Datenbank ist das Herzstück der täglichen Arbeit.

Hier werde alle für die Geschäftsprozesse anfallenden Daten erfasst und verwaltet.

Klassische OLTP-DB sind ERP, CRM u.a.

Diese Arte von DB ist auf schnelles und paralleles Erfassen von Daten im Multi-User Modus optimiert.

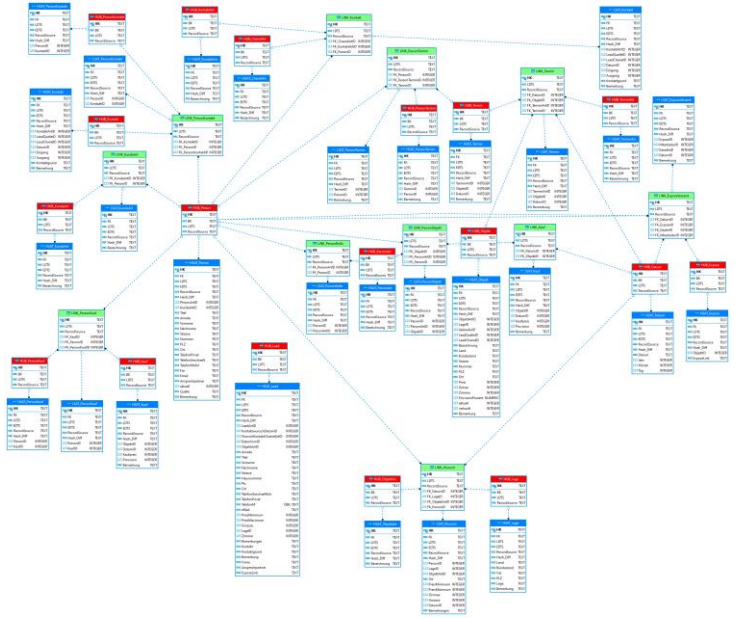


7.4 Data Warehouse

Das DWH ist das Herzstück für Reporting und Analyse von Daten. Es wird regelmäßig (täglich, stündlich, near real time) mit Daten aus der OLTP-DB und anderen Quellen befüllt (ergänzt).

Diese Art von DB ist designt und optimiert für komplexe Abfragen über große Datenmengen. OLTP-DB und DWH-DB haben sehr unterschiedlich technische Anforderungen und werden daher i.d.R. auf getrennten Systemen gehostet.

Wir verwenden ein Data Vault Modell aufgrund der vielen Vorteile: schnelle Entwicklung, einfache Wartung, Auditierbarkeit, Data Lineage, etc.

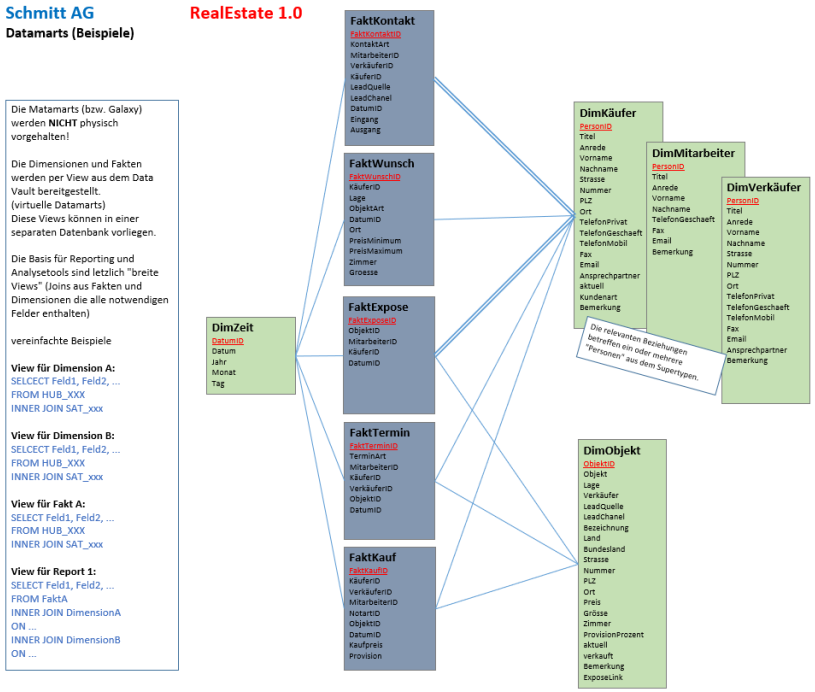


7.5 Data Mart(s)

Die DWH-Datenbank kann sehr groß und komplex werden und je nach innerer Struktur nicht für den unmittelbaren Konsum durch Gelegenheitsnutzer geeignet.

Daher werden oft Teile der DWH-DB separat aufbereitet und in kleinere, anders strukturierte DB übertragen.

Kriterien für diese „Auszüge“ können inhaltlicher, regionaler, rechtlicher Natur sein.



7.6 Reports und Analysen

Die Data Marts sind strukturell für das Erzeugen von Reports und Analysen optimiert.

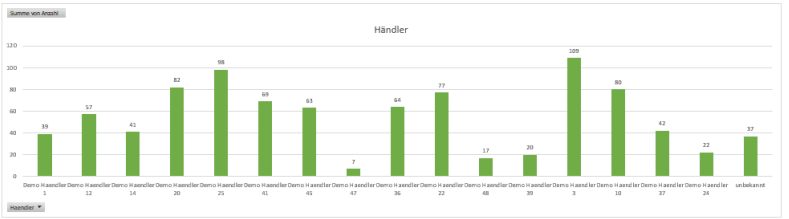
Sie liegen i.d.R. in Form einfacher Star- oder Galaxy-Schemas vor und erlauben auch unerfahrenen Nutzern einen schnellen und einfachen Zugang zu Informationen.

Es lassen sich so inhaltliche und zeitliche Zusammenhänge darstellen, wie etwa:

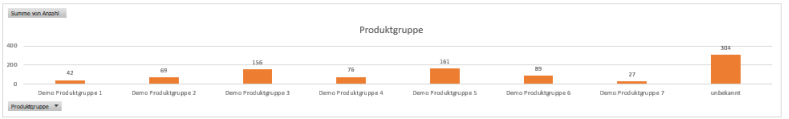
- Summen nach Produkten/Produktgruppen,
- Werte pro Jahr/Monat/Tag
- Veränderung von Werten über die Zeit
- Dauer bestimmter Prozesse.

Summe von Anzahl	Spaltenbeschrift								
Zeilenbeschriftungen	Demo Produktgrupp	Demo Produktgruppe 2	Demo Produktgruppe 3	Demo Produktgruppe 4	Demo Produktgruppe 5	Demo Produktgruppe 6	Demo Produktgruppe 7	unbekannt	Gesamtergebnis
Demo Haendler 1				19			20		39
Demo Haendler 12				57					57
Demo Haendler 14			41						41
Demo Haendler 20								82	82
Demo Haendler 25			98						98
Demo Haendler 41		69							69
Demo Haendler 45						63			63
Demo Haendler 47							7		7
Demo Haendler 36					64				64
Demo Haendler 22					77				77
Demo Haendler 48			17						17
Demo Haendler 39					20				20
Demo Haendler 3						13		96	109
Demo Haendler 10								80	80
Demo Haendler 37	42								42
Demo Haendler 24						13		9	22
unbekannt								37	37
Gesamtergebnis	42	69	156	76	161	89	27	304	924

Zeilenbeschriftungen	Summe von Anzahl
Demo Haendler 1	39
Demo Haendler 12	57
Demo Haendler 14	41
Demo Haendler 20	82
Demo Haendler 25	98
Demo Haendler 41	69
Demo Haendler 45	63
Demo Haendler 47	7
Demo Haendler 36	64
Demo Haendler 22	77
Demo Haendler 48	17
Demo Haendler 39	20
Demo Haendler 3	109
Demo Haendler 10	80
Demo Haendler 37	42
Demo Haendler 24	22
unbekannt	37
Gesamtergebnis	924



Zeilenbeschriftungen	Summe von Anzahl
Demo Produktgruppe 1	42
Demo Produktgruppe 2	69
Demo Produktgruppe 3	156
Demo Produktgruppe 4	76
Demo Produktgruppe 5	161
Demo Produktgruppe 6	89
Demo Produktgruppe 7	27
unbekannt	304
Gesamtergebnis	924



7.7 Live View

Viel besser als in PowerPoint Präsentationen lassen sich Details im lebenden System zeigen und erklären.

Daher ab ins Filesystem…

- 00----------
- 01-RealEstat_V1.png
- 02-Projekt.docx
- 03-Projektplan.xlsx
- 04-RealEstate_V1.pptx
- 10----------
- 11-machen-die-Kollegen.txt
- 20----------
- 22-OLTP-ERM.PNG
- 22-OLTP-ERM.xlsx
- 23-OLTP-Dictionary.csv
- 24-OLTP-NumberDemoRecords.csv
- 25-OLTP-Build.sql
- 30----------
- 31-DWH-ERM.png
- 33-DWH-Dictionary-HUB.csv
- 33-DWH-Dictionary-LINK.csv
- 34-DWH-Build.sql
- 40----------
- 41-Build.bat
- 42-sqlite3.exe
- 50----------
- 51-Datamart-ERM.PNG
- 52-Datamart-ERM.xlsx
- 59----------
- README.txt

Vielen Dank für ihre Aufmerksamkeit.

Wir freuen uns auf zukünftige Projekte.