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) **: ‘**3.11.20**2**11‘
- **Art des Objektes** (1FH, Mehrfamilienhaus, etc.) : Einfamilienhaus statt 1FH
- **Falsche Eintrag:** Expose zugesendet vor dem ersten persönlichen Telefonat

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
Unzulässiger Wert	Kauftermin = 18.02.2322	Werte außerhalb des Admin Bereichs
Formatfehler	Kaufpreis = 15.2**a**5.000 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

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
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

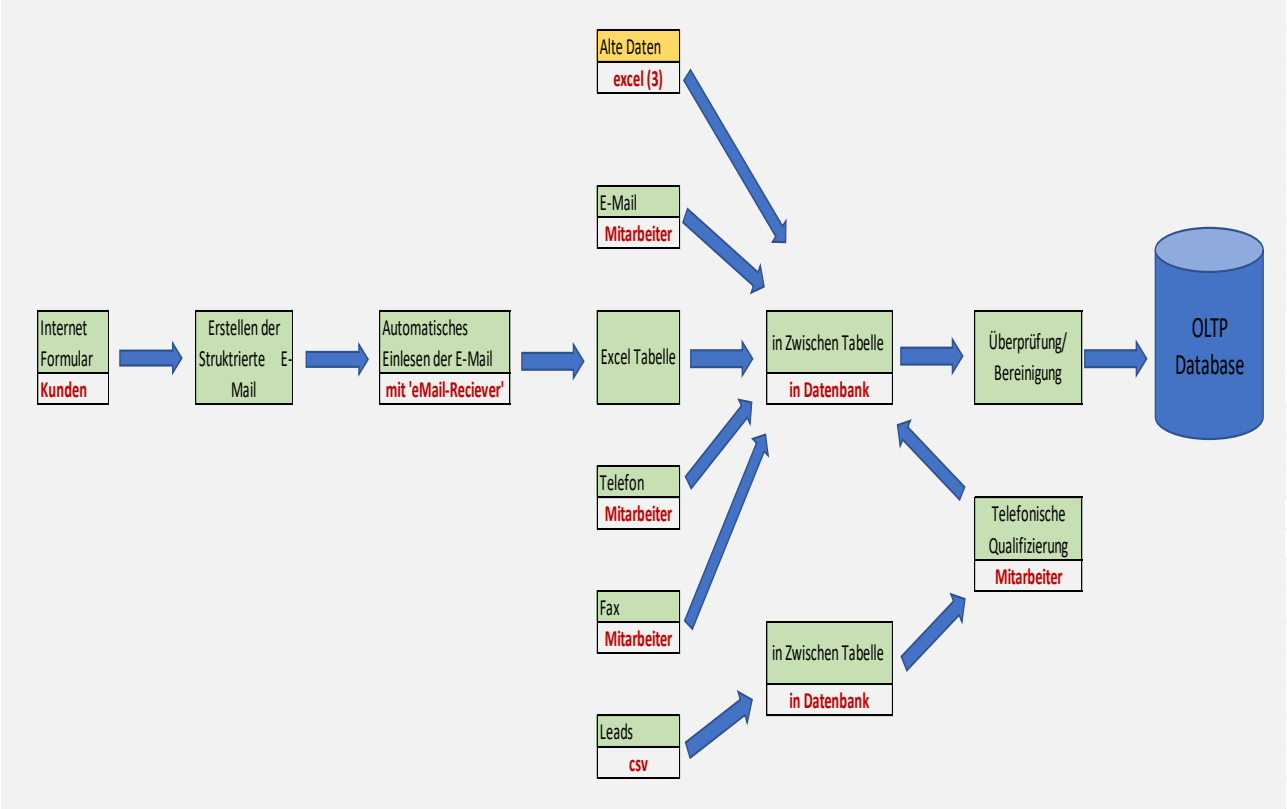
Duplikate:

Beispiel:

Um 8.00 Uhr ruft Herr Müller an und meldet sein Haus zum Verkauf. Um 16.00 Uhr ruft Frau Müller an und meldet das Haus nochmal zum Verkauf. Weil Sie nicht mit Ihrem Mann abgesprochen hat. Als Ergebnis würden in der Datenbank für dasselbe Objekt zwei Datensätze angelegt werden. Um das zu verhindern, könnte man folgenden Algorithmus integrieren:

Bilder Hashwert über die Spalten PLZ, Ort, Straße, Nummer und prüfe, ob dieser Hashwert bereits in der Datenbank verwendet wird. Ja/Nein: Die Entscheidung über das Verfahren mit diesem Datensatz ist fachlich zu klären. Weitere potenzielle Duplikate können nach ähnlichen Überlegungen behandelt werden.

Datenflüsse



Daten stammen aus unterschiedlichen Quellen und auf unterschiedliche Weise. Einige der möglichen Probleme, die während des Datenflusses auftreten können, sind wie folgt.

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 DLL	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