Data Engineering "ImmoMatch"

Präsentation Gruppe 2/B

Unsere Aufteilung Innerhalb der Gruppe

Aufgabe	**Hauptzuständige**	**Präsentation**	**Bemerkung**
Anforderungsdokumentation	Torben, Jürgen	Torben	
ERM	Natalia, Jürgen	Natalia	
Prototyp / DB Erstellung	Jürgen, Natalia	Jürgen	
DWH Modell	Daniel, Muhammed, Torben, Natalia	Daniel	
Data Quality	Muhammed, Daniel, Torben	Muhammed	

Verwendete Programme

- Draw.io
  - Kostenlos
  - Ansprechend
  - Vielseitig
- MySQL Workbench
  - Test einer professionelleren Lösung und Automatisierung, aber SQL
- SQLiteBrowser
  - Für jeden bekannt
- Google Drive: Docs, Präsentation
  - Sehr viel Zusammenarbeit

Kundenwünsche und Anforderungen

Der Kunde

- Weiteres Standbein wg. Covid-19
  - Immobilien
- Hauptziel: Matching
- Bis jetzt auf Excelbasis
  - Verkäufer, Objekte, Lage,...
  - Typische Datenbankoperationen werden später noch gezeigt
- Das Unternehmen wächst aktuell auf der Basis seiner 16 Mitarbeiter sehr schnell
  - Agiles und mächtigeres System nötig
  - 1000 Transaktionen täglich (10% Telefon, Rest Internetformular)
  - Nationaler Rahmen

Die Rahmenbedingungen

- Wichtige nicht-funktionalen Anforderungen:
  - Später zu schaffende GUI muss funktionieren
- Internetformular vorhanden
  - Übertragbarkeit
- Kunde ist noch in Aufbau
  - Detailierte Analysen nötig
- **Das Hauptziel unserer Arbeit ist es dem Kunden zu ermöglichen seine Ziele zu messen und zu evaluieren (Leads, Käufer, Verkäufer, Transaktionen,...), am Markt zu bleiben und den Umsatz zu steigern**

Typischen Aktionen für das operative System

- Abfragen nach einzelnen Attributen:
  - Kundennummer, Nachname, Objektid, Haustyp, Lage, Preismaximum, ...
- Abfragen nach Kombinationen:
  - Beispiele: Lage&Typ, Lage&Typ&Preis
  - Objekte eines Verkäufers
- Abfragen nach Auswertungen:
  - Besichtigungen pro Objekt
  - Abfrage Anzahl Exposés pro Kunde
  - Abfrage Anzahl versandter Exposés nach Objekt
  - Exposes(-kunden) eines Objekts, Exposes nach Kunden
- Abfragen mit vielen Attributen:
  - Mögliche Objekte für einen Kunden (Matching)
- Funktionen:
  - Mailing an alle Kunden, ...

Analyse-Anforderungen im Datawarehouse

- **Aktualität**: Tag
- **Historisierung** (z.B. Umsatz)
- Auswertung der **Leadqualität**
- **Analysen** DWH nach den folgenden Dimensionen (Beispiele)
    - **Art des Objektes**
      z.B. Welche Art von Objekt wurde vom Kunden am meisten nachgefragt?
    - **Lage**
      z.B. In welcher Lage besteht mehr oder weniger Nachfrage nach Objekten?
    - **Ort**
      z.B. Welche Stadt hat mehr oder weniger Nachfrage von Käufern oder Verkäufern?
    - **Preis**
      z.B. Wie der Preis des Objekts den Käufer beeinflusst?
    - **Zeit**
      z.B. Wie viele Besichtigungen gab es einem Quartal
    - **Effiktivität**
      z.B. Welcher Mitarbeiter verkauft am besten?, etc.
- Welche der erfassten Merkmale sind kaufentscheidend?

Gespräche mit dem Kunden

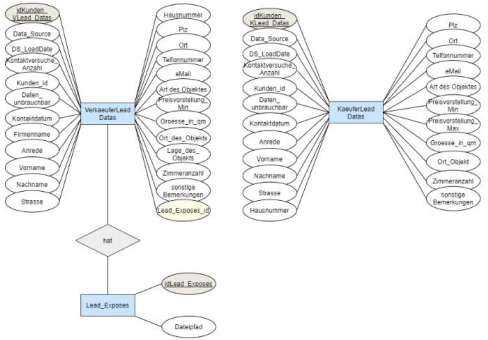
Gesprächstermine	Datum/Zeit	Bemerkung
“Gespräch 1”	2021-05-31 10:00	Sondierung der genaueren Entitätenstruktur und des Geschäftsmodells
“Gespräch 2”	2021-05-31 11:00	“Es gibt keine weiteren Kundendaten.”
“Gespräch 3”	2021-05-31 14:50	Mitarbeiter <-> Kunde Verkaufsprozess
“Gespräch 4”	2021-05-31 15:25	Eingekaufte Leads
“Gespräch 5”	2021-06-01 9:30	Leadvorqualifizierung
“Gespräch 6”	2021-06-01 12:26	Auswertung von Kontakten im DWH
“Gespräch 7”	2021-06-01 14:45	Measures im DWH: Provision

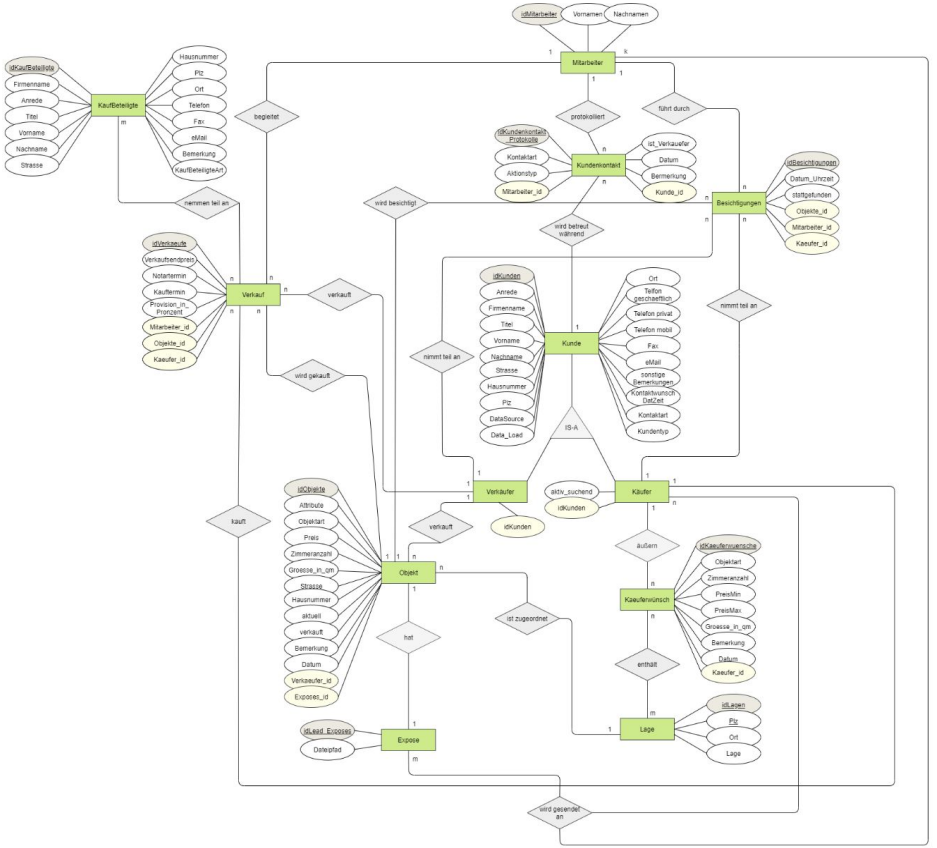
Best Practices

- Protokolle schreiben
- Wiederholung des eigenen Problem- / Lösungsverständnisses
- Originaldaten/Musterdaten zukommen lassen
- Unklarheiten hinterfragen
- Geschäftsprozesse immer wieder beschreiben lassen

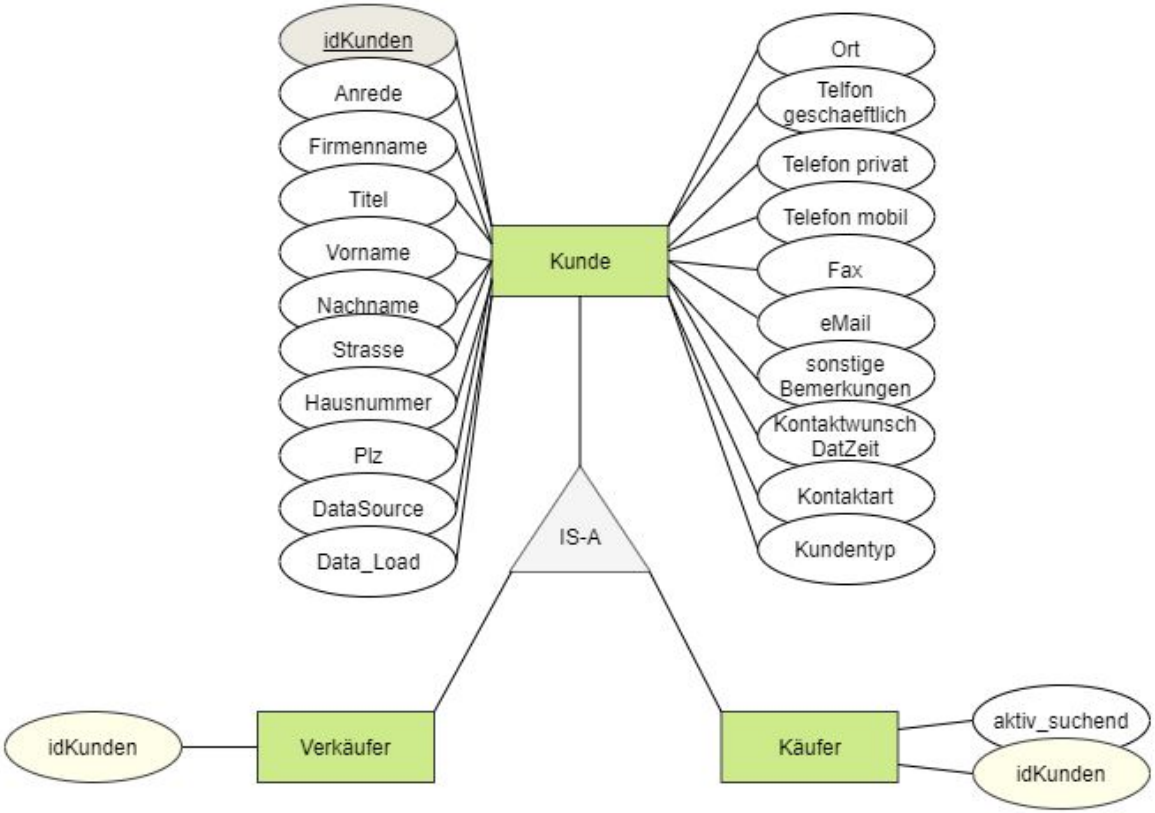
Natalia
ERM

Entity-Relationship Model (nach Chen, 1976)

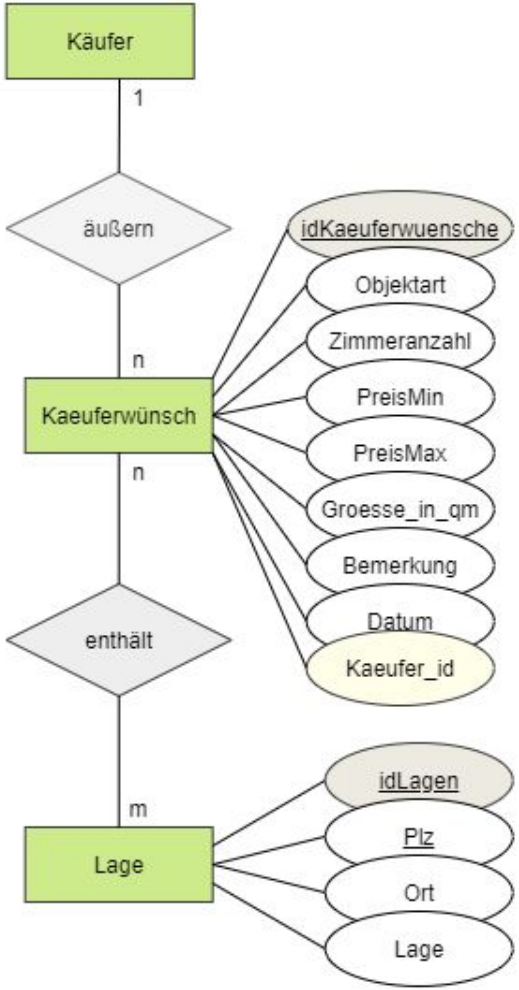




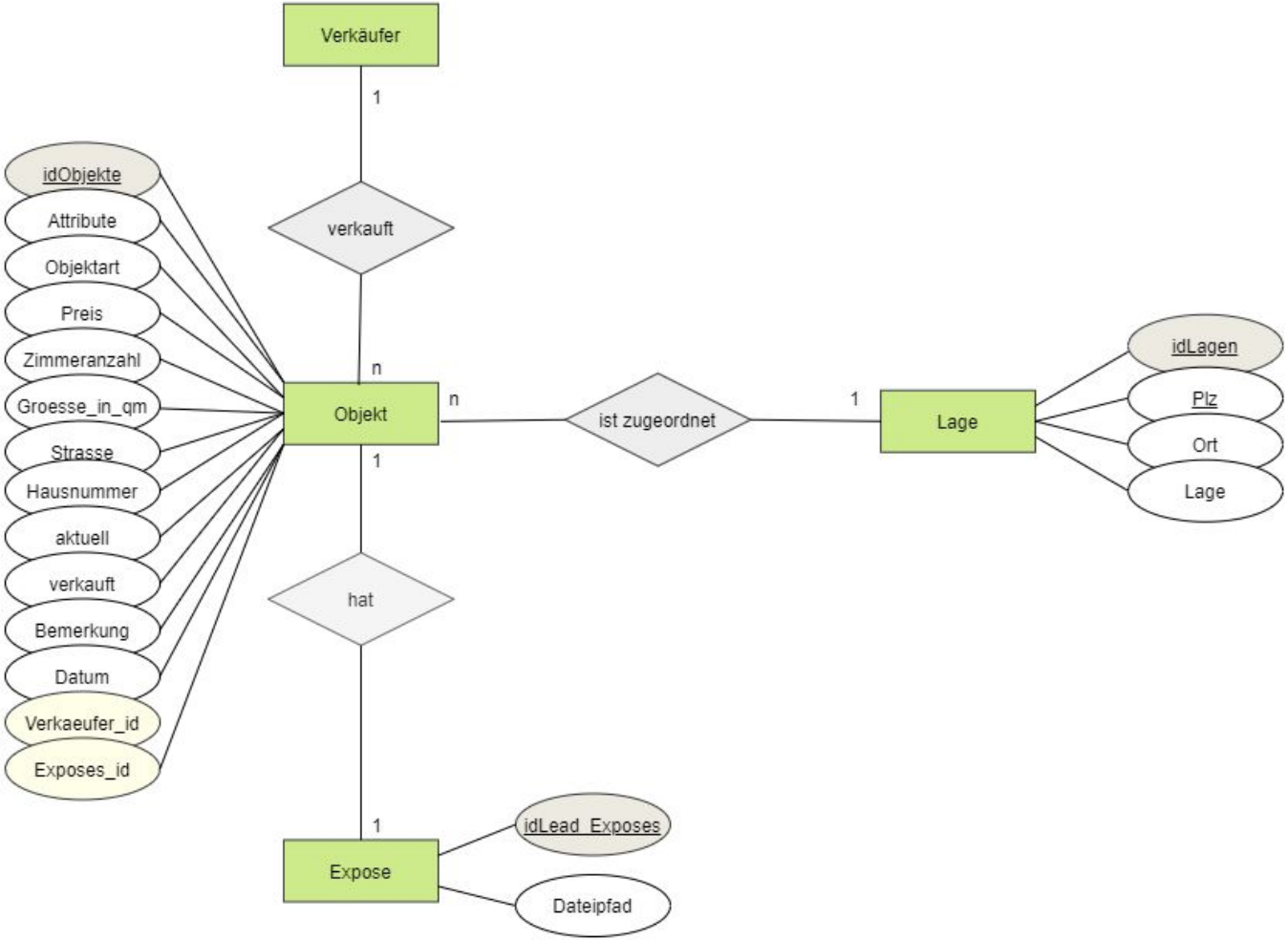
Kunden Abbildung



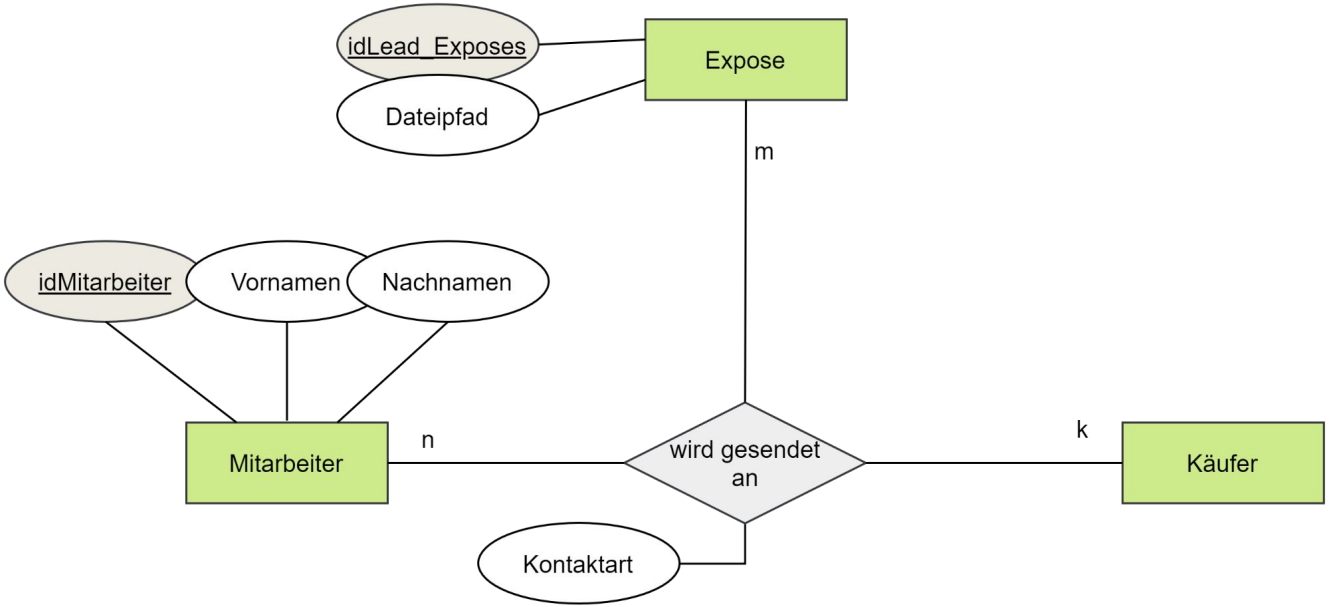
Kundenwünsche



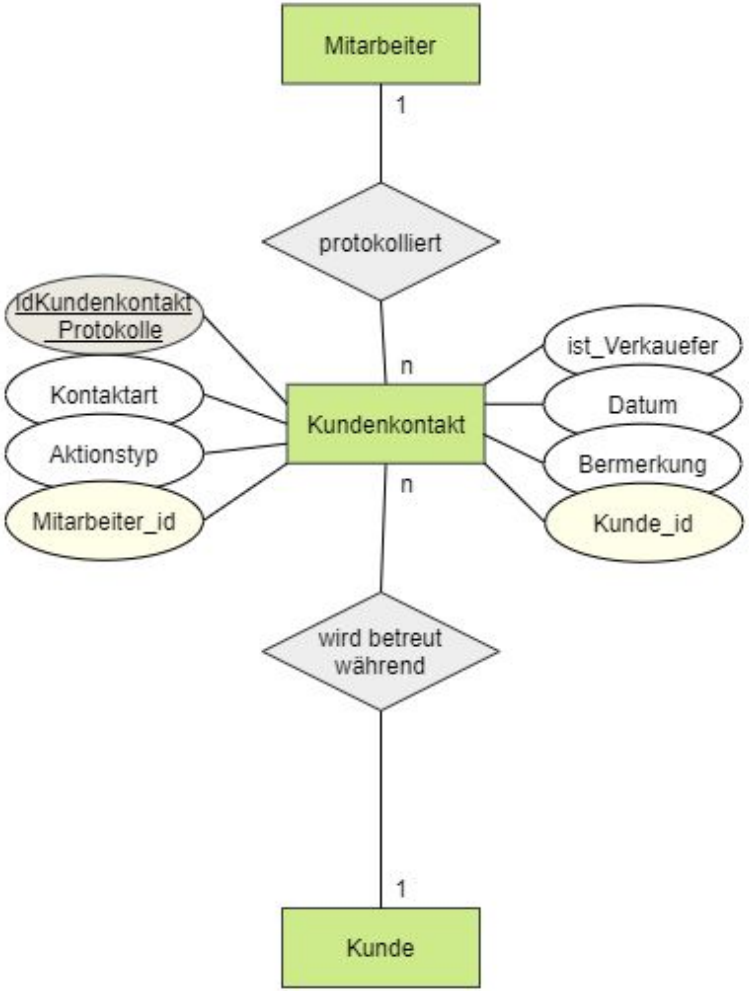
Verkaufsobjekt



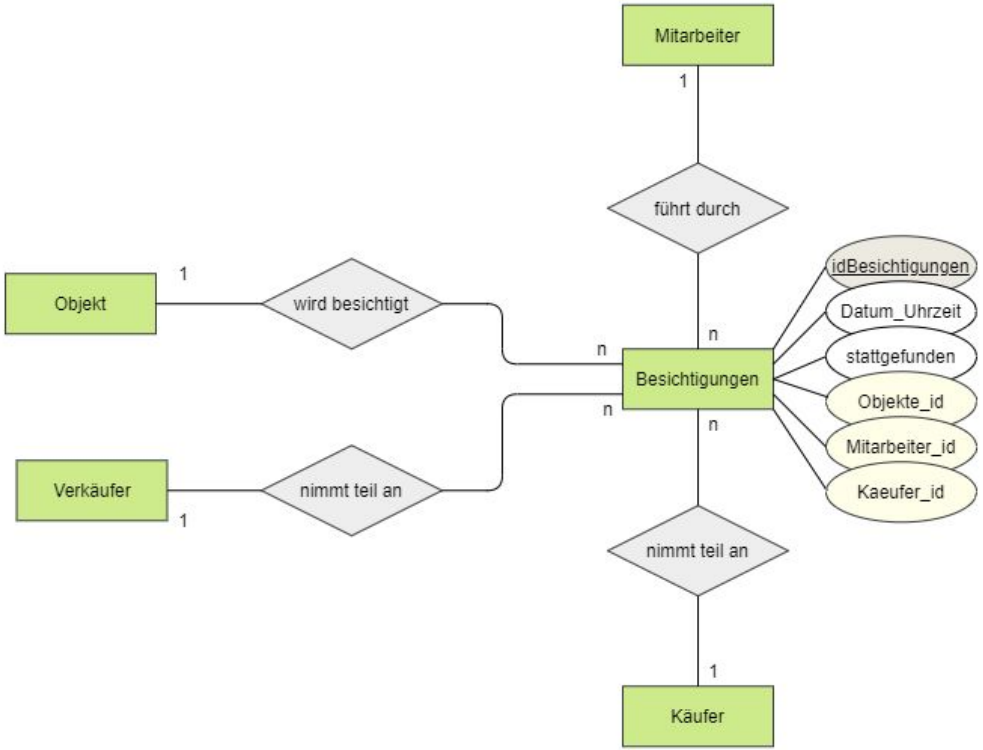
Exposé


idLead_Exposes
Dateipfad
Expose
m
idMitarbeiter
Vornamen
Nachnamen
Mitarbeiter
n
wird gesendet an
k
Käufer
Kontaktart

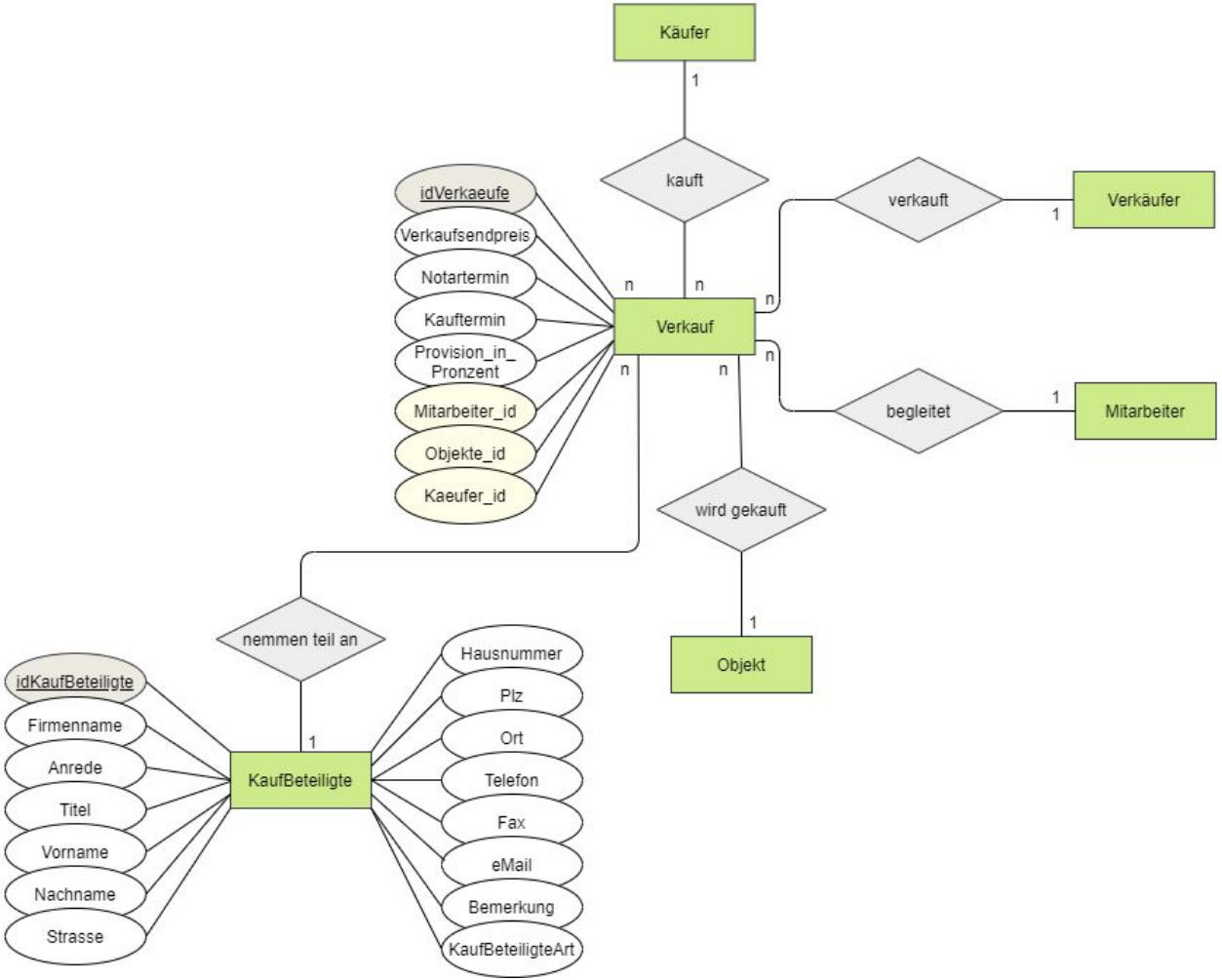
Kundenkontakt



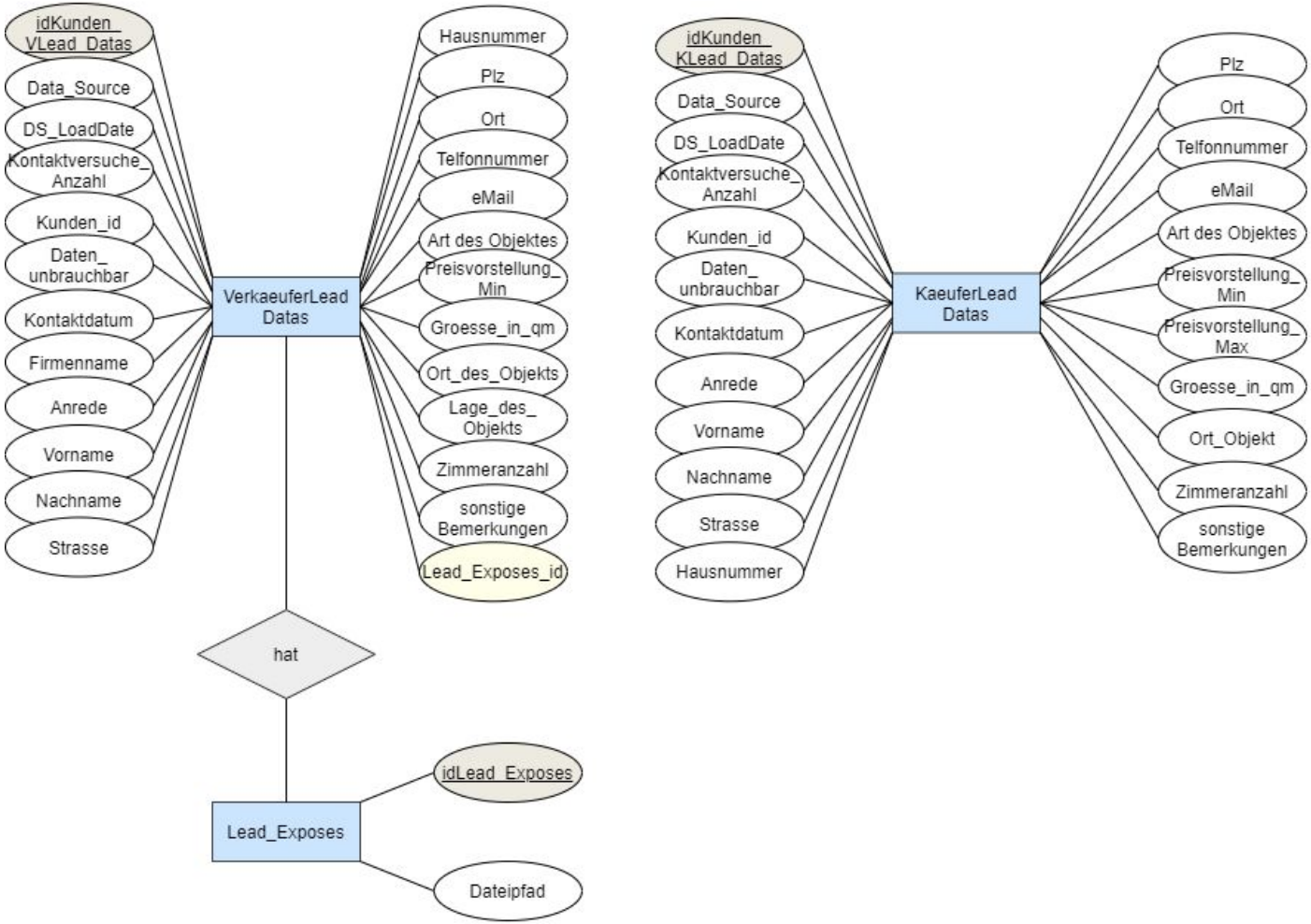
Besichtigungen



Kaufgeschäft

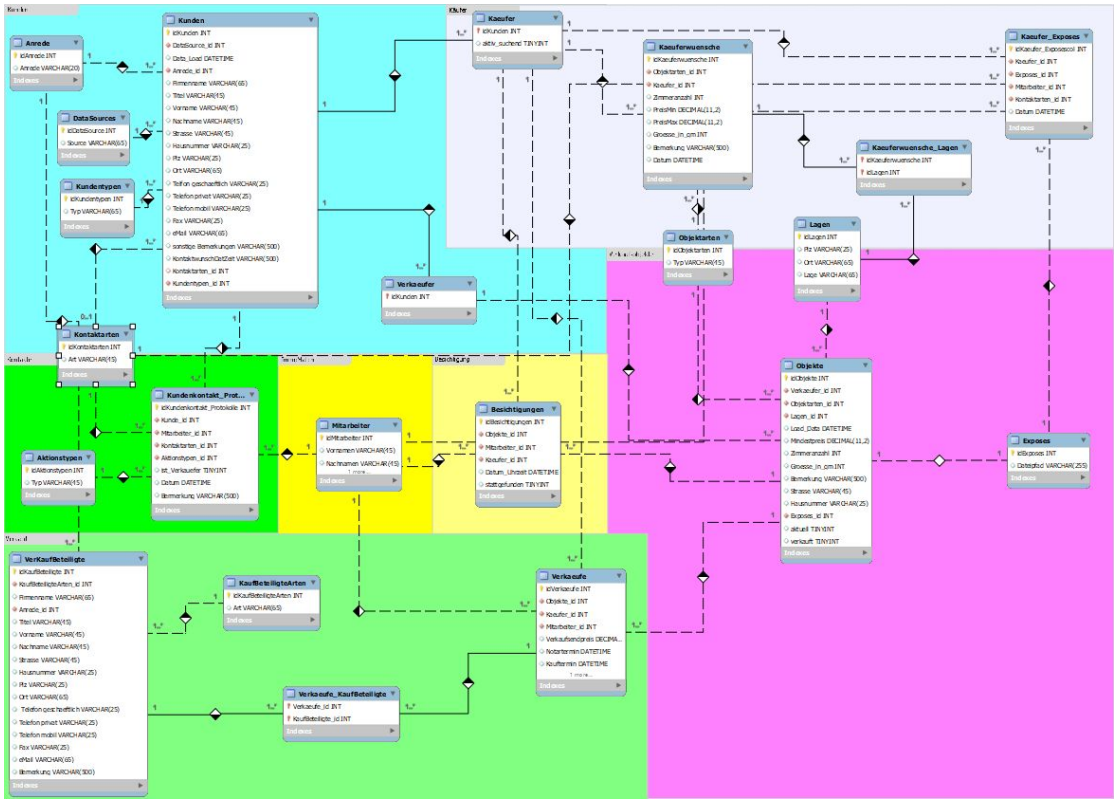


Leads Data

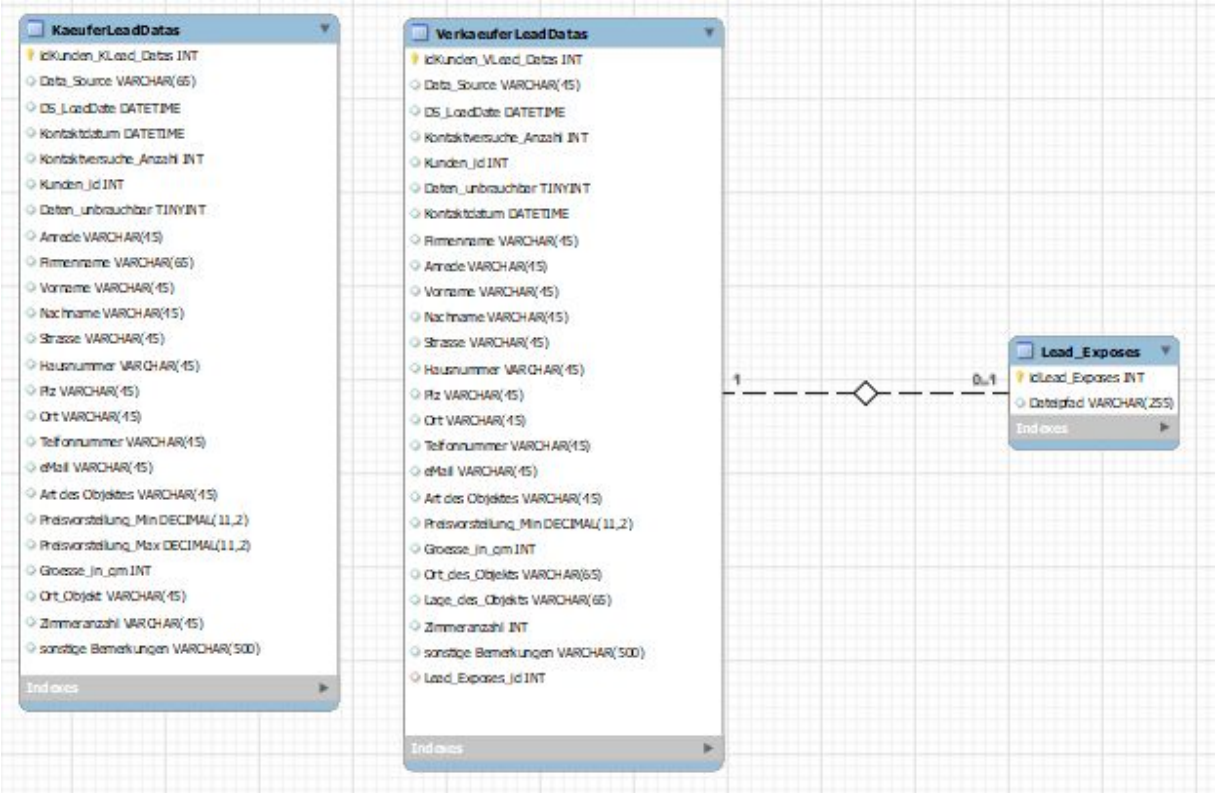

idKunden_ VLead_Datas
Data_Source
DS_LoadDate
Kontaktversuche_ Anzahl
Kunden_id
Daten_ unbrauchbar
Kontaktdatum
Firmenname
Anrede
Vorname
Nachname
Strasse
VerkaeuferLead Datas
Hausnummer
Plz
Ort
Telfonnummer
eMail
Art des Objektes
Preisvorstellung_ Min
Groesse_in_qm
Ort_des_Objekts
Lage_des_ Objekts
Zimmeranzahl
sonstige Bemerkungen
Lead_Exposes_id
hat
idLead_Exposes
Lead_Exposes
Dateipfad
idKunden_ KLead_Datas
Data_Source
DS_LoadDate
Kontaktversuche_ Anzahl
Kunden_id
Daten_ unbrauchbar
Kontaktdatum
Anrede
Vorname
Nachname
Strasse
Hausnummer
KaeuferLead Datas
Plz
Ort
Telfonnummer
eMail
Art des Objektes
Preisvorstellung_ Min
Preisvorstellung_ Max
Groesse_in_qm
Ort_Objekt
Zimmeranzahl
sonstige Bemerkungen

Jürgen
Prototyp DB

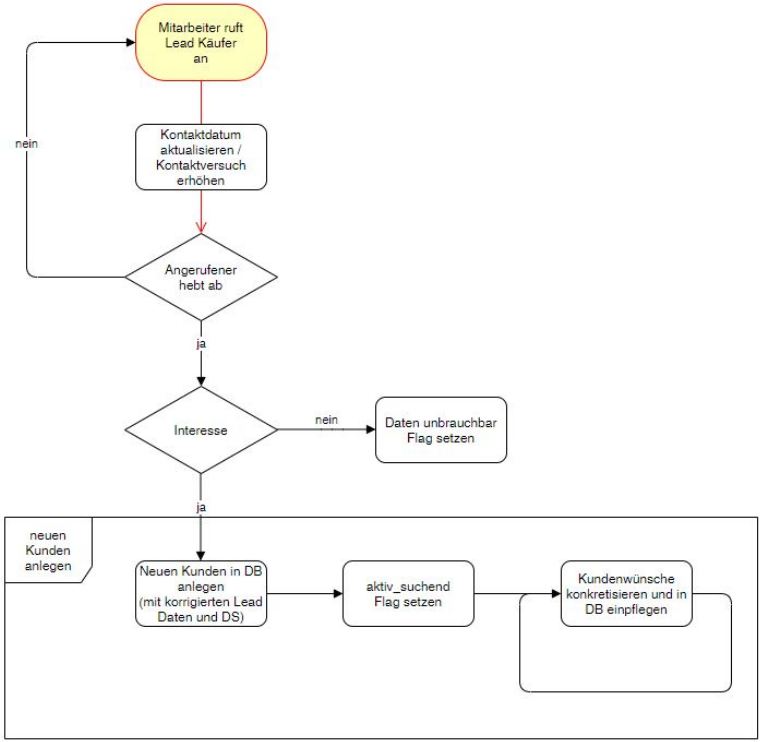
DB-Model in MySQL Workbench



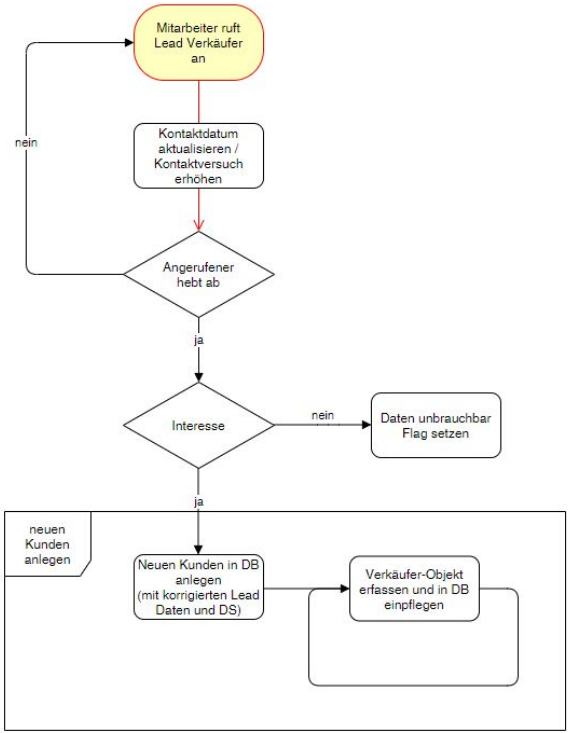
DB-Model in MySQL Workbench



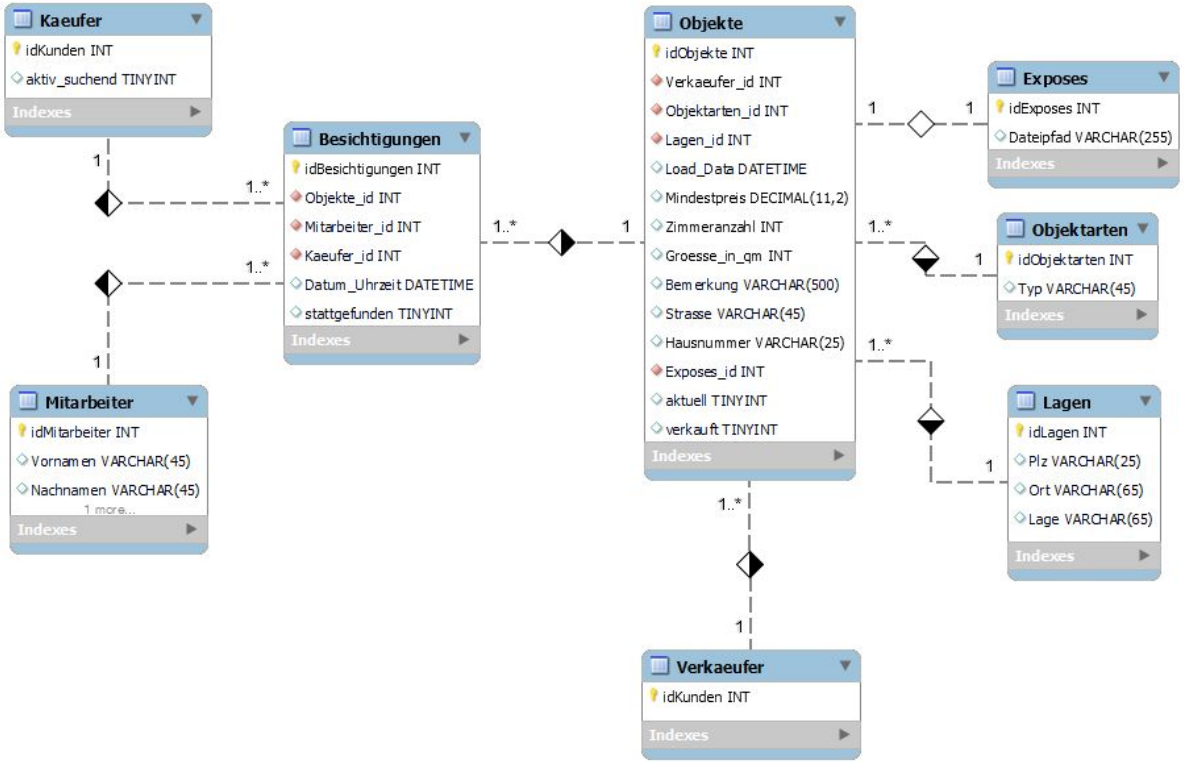
Käufer-Lead kontaktieren



Käufer-Lead kontaktieren



Beispiel Abfragen für Objekte 1

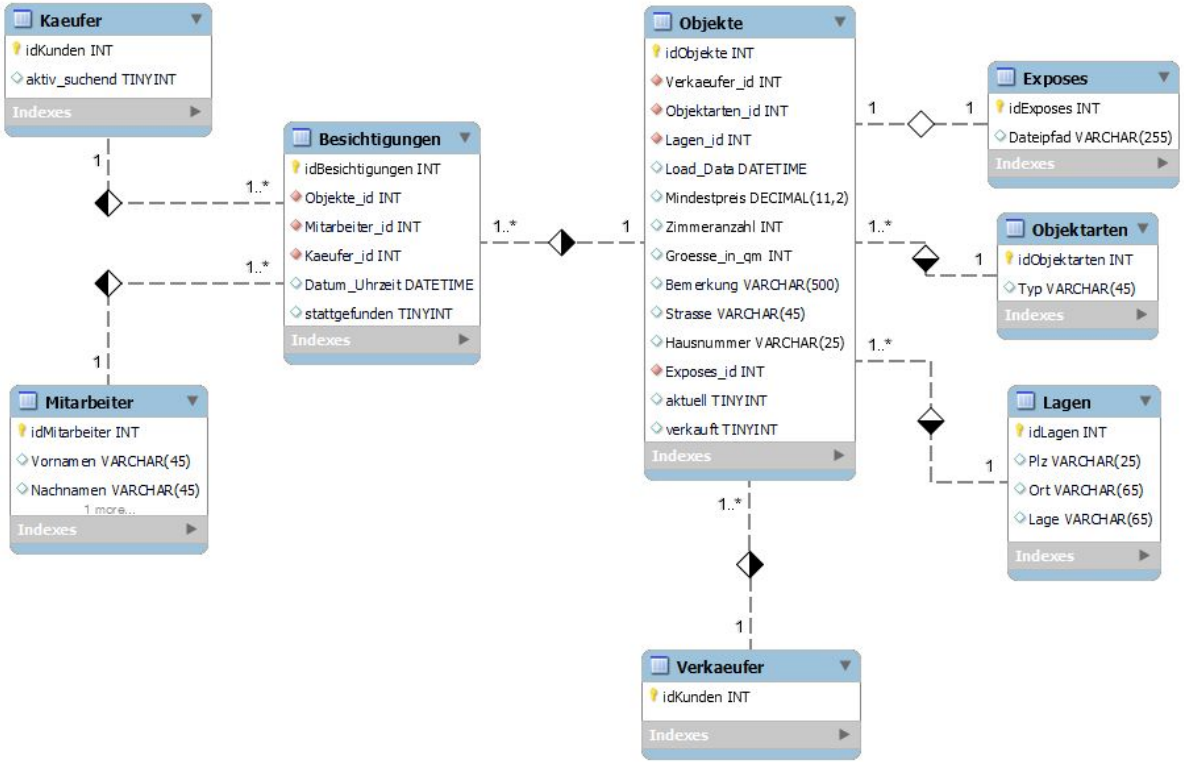


Anzahl der Besichtigungen pro Objekt

Pseudo-SQL:

Zähle alle Eintrage in Besichtigung mit gleicher Objekt_id

Beispiel Abfragen für Objekte 2

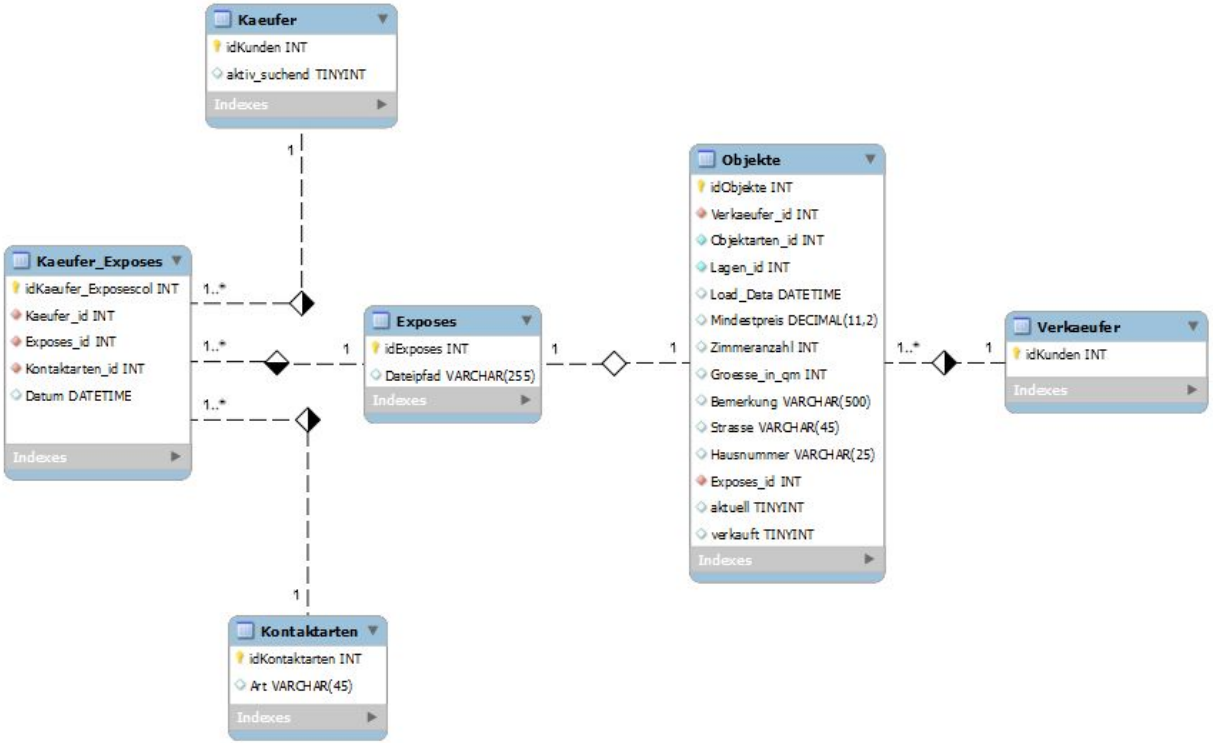


Alle Besichtigungen des Mitarbeiter x in der Lage y

Pseudo-SQL:

Filter Besichtigungen nach Kaufer_id und Objekt_id (Objekt_id = Alle Objekte mit entsprechender Lagen_id in Objekte)

Beispiel Abfragen für Expose 1

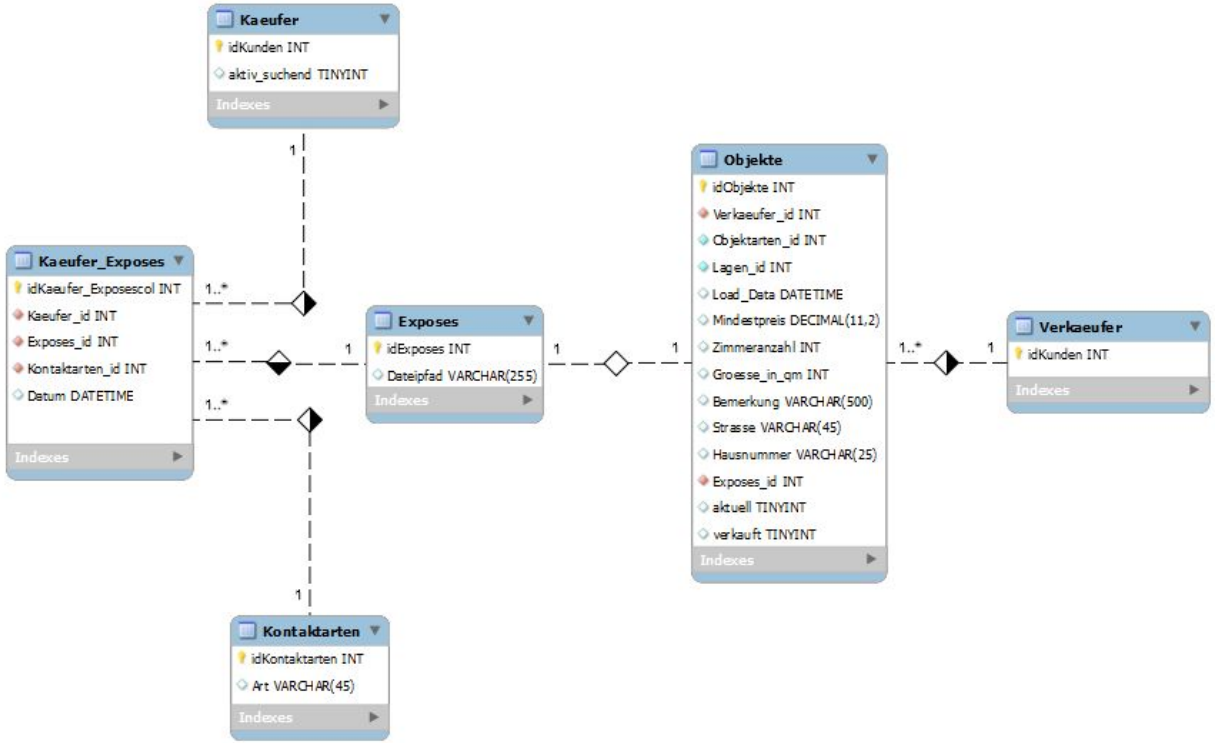


Exposes pro Kunde:

Pseudo-SQL:

Zähle in Kaeufer_Exposes die Anzahl der Exposes pro Kaeufer_id

Beispiel Abfragen für Expose 2



Alle Kaufer die Exposes zugeschickt bekommen haben und deren Objekt-Mindestpreis über 900.000 Euro liegt:

Pseudo-SQL:

Finde alle Objekte mit Mindestpreis > 900.000 Euro

Lese die zugehörigen Exposes Ids aus

Finde in Kaeufer_Exposes alle Kaeufer mit den gefunden Exposes_id

Lokalisiere daraus die entsprechenden Käufer

Daniel
DWHM

Darstellung DWH ⟹ Das Galaxy schema

Warum?

- Aufgrund der typische Aktionen vom Abteilungsleiter der Real Estate geht das Schema auf Ihre Bedürfnisse ein.
- Dieses komplexes, logisches Architekturkonzept besitzt ein gutes Abfrageverhalten.
- Die gewünschten Abfragen werden möglich gemacht werden.

Darstellung DWH

Fact Tabelle:

- Kontakt_Fakt
- Verkaeufe_Fakt

Kontakt_Fakt	
PK, FK	Mitarbeiter_Dim_id
PK, FK	Verkaeufer_Dim_id
PK, FK	Objekte_Dim_id
PK, FK	Kaeufer_Dim_id
PK, FK	Kontaktsdatum
	Kontaktsart
	Aktionstyp

Verkaeufe_Fakt	
PK, FK	Verkaufsdatum
PK, FK	Objekte_Dim_id
PK, FK	Kaeufer_Dim_id
PK, FK	Verkaeufer_Dim_id
PK, FK	Mitarbeiter_Dim_id
	Verkaeufsendpreis
	Provision

Dimensionale Tabelle

- Mitarbeiter_Dim
- Kaeufer_Dim
- Datum_Dim
- Objkete_Dim
- Verkaeufer_Dim

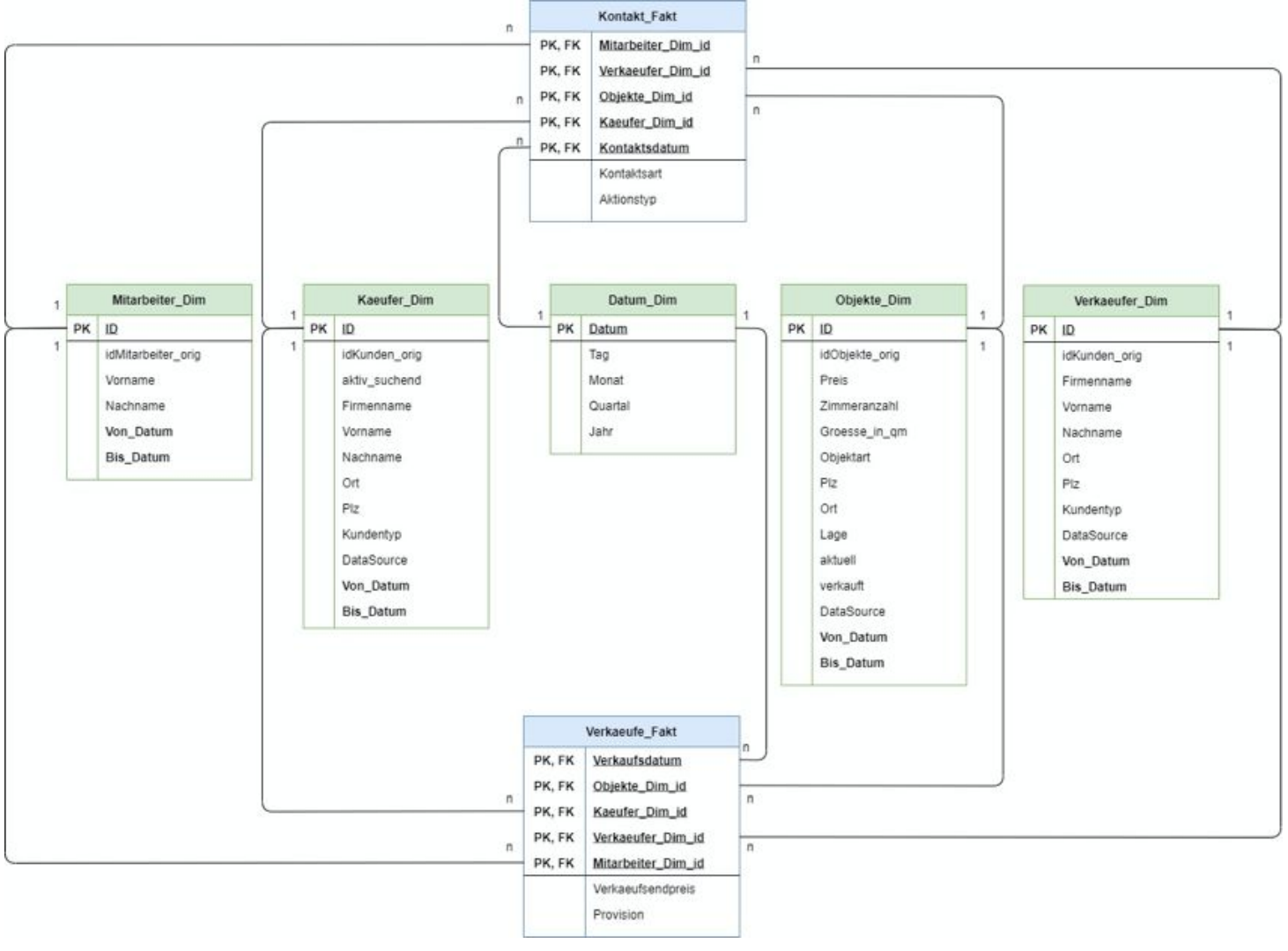
Mitarbeiter_Dim	
PK	**ID**
	idMitarbeiter_orig
	Vorname
	Nachname
	Von_Datum
	Bis_Datum

Kaeufer_Dim	
PK	**ID**
	idKunden_orig
	aktiv_suchend
	Firmenname
	Vorname
	Nachname
	Ort
	Plz
	Kundentyp
	DataSource
	Von_Datum
	Bis_Datum

Datum_Dim	
PK	**Datum**
	Tag
	Monat
	Quartal
	Jahr

Objekte_Dim	
PK	**ID**
	idObjekte_orig
	Preis
	Zimmeranzahl
	Groesse_in_qm
	Objektart
	Plz
	Ort
	Lage
	aktuell
	verkauft
	DataSource
	Von_Datum
	Bis_Datum

Verkaeufer_Dim	
PK	**ID**
	idKunden_orig
	Firmenname
	Vorname
	Nachname
	Ort
	Plz
	Kundentyp
	DataSource
	Von_Datum
	Bis_Datum


Kontakt_Fakt
PK, FK Mitarbeiter_Dim_id
PK, FK Verkaeufer_Dim_id
PK, FK Objekte_Dim_id
PK, FK Kaeufer_Dim_id
PK, FK Kontaktsdatum
Kontaktsart
Aktionstyp
Mitarbeiter_Dim
PK ID
idMitarbeiter_orig
Vorname
Nachname
Von_Datum
Bis_Datum
Kaeufer_Dim
PK ID
idKunden_orig
aktiv_suchend
Firmenname
Vorname
Nachname
Ort
Plz
Kundentyp
DataSource
Von_Datum
Bis_Datum
Datum_Dim
PK Datum
Tag
Monat
Quartal
Jahr
Objekte_Dim
PK ID
idObjekte_orig
Preis
Zimmeranzahl
Groesse_in_qm
Objektart
Plz
Ort
Lage
aktuell
verkauft
DataSource
Von_Datum
Bis_Datum
Verkaeufer_Dim
PK ID
idKunden_orig
Firmenname
Vorname
Nachname
Ort
Plz
Kundentyp
DataSource
Von_Datum
Bis_Datum
Verkaeufe_Fakt
PK, FK Verkaufsdatum
PK, FK Objekte_Dim_id
PK, FK Kaeufer_Dim_id
PK, FK Verkaeufer_Dim_id
PK, FK Mitarbeiter_Dim_id
Verkaeufsendpreis
Provision

Mitarbeiter_Dim_ID	Kaeufer_Dim_id	Verkaeufer_Dim_ID	Objekte_Dim_ID	Kontaktdatum+Uhrzeit	Kontaktart	Aktionstyp
99999999	1113 1	99999999	99999999	20.05.21 2	Internetformular 3	Kundenerstellung 4
99999999	1113	99999999	2905	20.05.21	Internetformular	Objekterzeugung
11	217	99999999	99999999	21.05.21	Email	Exposees angefragt
11 1	217 2	19 3	23	21.05.21 4	Email 5	Exposees gesendet 6
11	217	305	527	21.05.21	Email	Exposees gesendet
11	217	19	23	23.05.21	Telefon	Besichtigungstermin vereinbart
11 1	217 2	19 3	23	27.05.21 4	Direkter Kontakt	Besichtigungstermin durchgeführt 5

Ist_Verkaeufer

Über ExposeId

Über Objekt

Kundenkontakt_Protokolle

- IdKundenKontakt_Protokolle (PK)
- Kunde_id (FK)
- Mitarbeiter_id (FK) 1
- Kontaktarten_id (FK) 3
- Aktionstyp_id (FK) 4
- ist_Verkaeufer
- Datum 2
- Bemerkung

Besichtigungen 5

- idBesichtigungen (PK)
- Objekte_id (FK) 3
- Kaeufer_id (FK) 2
- Mitarbeiter_id (FK) 1
- Datum_Uhrzeit 4
- stattgefunden

Kaeufer_Exposes 6

- idKaeufer_Expose (PK)
- Kaeufer_id (FK) 2
- Exposes_id (FK) 3
- Mitarbeiter_id (FK) 1
- Kontaktart_id (FK) 5
- Datum 4

DWH - historisierung zur Weiteren Analyse

- Nachfassen bei alten Kunden
- Umsatz pro Tag, Monat, Jahr
- Kundenliste von Objekt
- Die Tage oder Monate, in denen es mehr Käufe gibt
- Für Sicherung der Daten und als Leads
- Tagesaktuell ausreichend

Muhammed DQ

Potenzielles Problem

Scope	Dirty Data	Reasons
Unzulässiger Wert	datum = 30.02.2050	Werte außerhalb des Admin Bereichs
Attributabhangigkeit verletzt	Notartermin > Kauftermin	Kommt je nachdem
Eindeutigkeit verletzt	(Ort = Berlin, PLZ = 123) (Ort =Mannheim, PLZ = 123)	Uniqueness für PLZ
Referentielle Integritat verletzt	name = “Peter Müller”, idObjekte = “123”	referenzierte Objekte (123) nicht definiert
Fehlende Werte	Telefonnummer = +(49)15….	falsches Format oder zusätzliche Informationen
Schreibfehler	Preis = 23.49 oder 23,49	falsche Symbole
Kryptische Werte, Abkürzung	title = Doktor title = Dr.	verschiedene Einträge
Eingebettete Werte	Name = “Georges Spaninen”	mehrere Werte in einem Attribut
Falsche Zuordnung	Stadt = “Spanien” oder “Spain”	mehrere Sprachen
Widersprüchliche Werte	Stadt = “Berlin”, PLZ=33333	Stadt und PLZ müssen abhängig sein
Transpositionen	Name = “Torben E.” Firma = “Bosch T”	normalerweise in einem Freiformfeld
Duplikate	kunden1(name = “Peter Müller”, ...) kunden2(name = “P. Müller”, ...)	gleiche Kunden aufgrund einiger Dateneingabe Fehler zweimal vertreten oder Frau und Mann trägt unabhängig voneinander ein
Datenkonflikte	kunden1(name = “Peter Müller”, Kauftermin = 12.06.2021) kunden2(name = “Peter Müller”, Kauftermin = 17.06.2021)	dieselbe reale Entität wird durch unterschiedliche Werte beschrieben

Probleme	Data	Ansatz
Unzulässiger Wert	Kardinalität	Falls Kardinalität > Erwartung >>>> Problem
Unzulässiger Wert	max, min	max, min sollte nicht außerhalb des zulässigen Bereichs liegen
Unzulässiger Wert	var, sd	Varianz, Abweichung der statistischen Werte sollte nicht höher als der Schwellenwert sein
Schreibfehler	Attribute	Beim Sortieren nach Werten werden oft falsch geschriebene Werte neben korrekten Werten angezeigt
Fehlende Werte	Nullwerte	Prozentsatz/Anzahl der Nullwerte
Fehlende Werte	Standardwerte	Die Standardwert kann darauf hinweisen, dass der tatsächliche Wert fehlt
Falsche Zuordnung	Attribute	Vergleichen des Attributwertesatzes einer Spalte einer Tabelle mit referenz Tabelle
Duplikate	Kardinalität	Einzigartigkeit von Kardinalität
Duplikate	Attribute	Sortieren von Werten nach der Häufigkeit des Auftretens; mehr als 1 Vorkommen weist auf Duplikate hin

- Definieren von Datentransformationen

- geeigneten Sprache

- Es ist notwendig, die erforderlichen Transformationen in einer geeigneten Sprache anzugeben,
    - Verschiedene ETL-Tools bieten diese Funktionalität, indem sie proprietäre Regelsprachen unterstützen, Oracle Warehouse Builder (OWB), SAP Data Services. IBM Infosphere Information Server. SAS Data Management, SQL Server Integration Services (SSIS)
    - Ein allgemeinerer und flexiblerer Ansatz ist die Verwendung der Standard Abfragesprache SQL, um die Datentransformationen durchzuführen und die Möglichkeit anwendungsspezifischer Spracherweiterungen zu nutzen

- **Attribute Splitting**
  - Eingebettete Werte
  - Name
- **Validierung und Korrektur**
  - Falsche Zuordnung, Unzulässiger Wert, Schreibfehler, Kryptische Werte, Abkürzung
  - Adresse
- **Standardisierung**
  - Unzulässiger Wert, Schreibfehler, Kryptische Werte, Abkürzung
  - Datum

Möglichkeiten zur Vermeidung von Fehler

Fehlervermeidungsstrategie	Vermeidung von….
DropDown Menüs	Unzulässiger Wert, Schreibfehler, kryptische Werte, Abkürzungen, falsche Zuordnung, Ort, Anrede
Pflichtfelder	Fehlende Werte, Kontaktdaten
einige Einschränkung definieren	Schreibfehler, Telefonnummer
Beispiel eines gewünschten Datensatz z.B. in grau hinterlegt	Unzulässiger Wert, Falsche Zuordnung Email Adresse, Telefonnummer
Korrekte Datensätze verwenden, einkaufen	falschen Daten

Tabellen Datenanalyse

- Fehlende Primärschlüssel (bisher nicht vorhanden)
- Vorname, PLZ: fehlende Daten (z.B. Frau Juliane Bilger hat keine PLZ)
- Vorname: Schreibfehler, Abkürzung (z. B Frau Aydemir hat Ihre Name als P. geschrieben)
- Ansprechpartner: fehlende Daten, Datenintegrität (die meisten sind leer)
- Straße: Formatfehler (z.B. Frau Juliane Bilger Straße: “Albert”)
- Falls Firma, gibt es Problem mit Vor und Nachname
- Anschprechpartner Partner und Email sind leer
- Es gibt zwei Spalte -> Ort (der Ort für Kunden und Objekt als Ort benannt)
- Für Tel FN, ein entry ist weniger als anderen, deswegen kann es so sein, dass User nur 9 zahlen eingeben soll
- Preis und Größe in qm sollen Ausreißer Daten überprüft werden
- Duplikate, Schreibfehler können logisch überprüft werden (z.B Sortieren von Werten wie Nachname, Ort)

Ausblick, Zusammenfassung, Best Practices & Learnings

Ausblick auf zukünftige Schritte

- Einpflegen aller bestehenden Daten
- Andere Währung wie CHF vorsehen
- Einführung einer Prozessnummer für bessere Nachverfolgbarkeit/Analyse
- Verfeinerung der Datentypen (nach Rücksprache mit Kunden)
- Aufbau einer GUI
- UML Modellierung
- Testphase
- Rollout
    - Schulung
    - Dokumentationen
- Rückmeldung / Verbesserung
- Mitarbeiterdaten ausbauen, hinzufügen, vereinheitlichen

Best Practice

Echtes Lasten-/Pflichtenheft wäre sehr hilfreich, aber in diesem zeitlichen Umfang leider nicht machbar