Management Datenqualität

Datenqualität in unserem Datenbank-System

Warum ist Datenqualität wichtig?

Warum ist Datenqualität in unseren Systemen wichtig?

- Sicherstellung von aussagekräftigen Datenanalysen („garbage in, garbage out“)
- Erkennung und Behebung von Inkonsistenzen, Fehlern, Informationsdefiziten in Daten

Beispiele was passieren kann:

- Objekte werden doppelt geführt und Verkäufer doppelt angeschrieben
- Die Matching-Ergebnisse sind falsch
- Wir führen Verkaufsgespräche zu bereits verkauften Objekten durch (sog. „False Postives“)
- Wir führen parallele Verkaufsgespräche ohne dies zu wissen
- Wir haben wertvolle Objekte im Portfolie und wissen es nicht (sog. „False Negatives“)

Warum ETL und Datenqualitätsprozess?

- Konsolidierte Sicht lässt erst schlechte Datenqualität erkennen
- Im Date Warehouse akkumulieren sich die Probleme
- Teure strategische Entscheidungen werden falsch getroffen

Beispiele Fehlerarten:

- Falsche Primär- und Sekundärschlüssel führen dazu, das falsche Daten kombiniert werden
- Fehlende Daten
- Falsche Daten – entweder objektiv (z.B. falsche Preise, Postleitzahlen, etc) oder widersprüchliche Werte
- Formatfehler – z.B. im Datum, Adressformat führen zu Fehlern
- Duplikate
- Schlechte/fehlende Dokumentation

Wie können Fehler passieren?

- Fehlerhafte Dateneingabe oder Erfassung
- Veralterung, z.B. Umzüge
- Fehlerhafte Transformation und Integration

Datenbereinigungskonzept

Was können wir gegen schlechte Datenqualität tun?

Datenqualitätskriterien bestimmen

- Definition von Bedingungen, denen die Daten entsprechen sollen, z.B. Mindestquadratmetergröße

Fehler bei Entstehung vermeiden

- Prüfungen auf Nutzereingaben bei Internetformulare → wird bereits so praktiziert
- Geschulte Mitarbeiter bzgl. Mitarbeitereingaben → wird bereits so praktiziert

Profiling

- Profiling-Erkunden des Datenbestandes - Manuell & Toolbasiert (Data Profiler)

Rollen und Logik gestützte systemische Prozesse

- Kontroll-System ETL Prozess & Datenqualitätsprozess

Monitoring

- Monitoring –Maßnahmen zur Fehlerbeseitigung evaluieren und kontrollieren

Spezielle Themen bei zugekauften Leads

Dubletten

Folgende Problematik ist aufgetreten. Die sehr teuer eingekauften Leads weisen sind häufig Dubletten von bereits in der Datenbank gespeicherten Objekten und Interessenten.

Lösung Hash-Funktion: Mit Hilfe der Hash-Funktion können Hash-Werte kalkuliert werden. Diese aggregieren die Inhalte eines Datensatzes zu einem Hashwert. Die Hashwerte unserer Datenbank können den Lead-Agenturen zur Verfügung gestellt werden. Die Agenturen dürfen uns keine Datensätze mit denselben Hashwerten zur Verfügung stellen.

Falsche oder unpassende Formate

Vorgaben an die Lead-Agenturen, wie die Daten in den SV Dateien gespeichert werden sollen, z.B. bzgl. Adressformate, Telefonnummern, Datumsformate, Sprache, …