Prof. Dr.-Ing. Thomas Kudraß
HTWK Leipzig, F IMN

Übungsblatt 5: Extraktion, Transformation, Laden (ETL)

1. Datenaufbereitung

Beschreiben Sie die Phasen der Datenaufbereitung im Data Warehouse.

2. Datenfehler

Welche Datenfehler können in einer operativen Quelle oder in mehreren Quellen bestehen? Wie kann man diese finden? Welche Fehler bleiben bei welchem Verfahren unberücksichtigt?

3. Datenqualität

Welche Datenqualitätsfehler sind in den Relationen *Kunde* und *Bestellung* zu finden? Durch welche Analysearten können die Fehler identifiziert werden?

Kunde

KNr	Nachname	Vorname	Adresse	Stadt	Geburtstag
555666	Maier	Thomas	First Avenue 12	New York	1983-10-10
123456	Muster	Max	Rue du Tour	Lyon	1972-01-01
112233	Schulz	Maik	M.-Gorki-Str. 5	Magdeburg	1990-12-03
445566	Thomas	Maier	Rue du Gare 11	Paris	NULL
123456	Schulz	Mike	Maxim-Gorki-Str. 5		1985-08-08

Bestellung

BNr	KNr	Artikel	Menge	Zugestellt
125	555666	4123649700201	1	T
512	123456	4222451689005	Zwei	1
699	112233	40815487990	3	0
730	555566	4900043174599	6	Nein
938	123456	3900004433901	Eins	Ja

4. Ähnlichkeitsbestimmung I

Berechnen Sie die Edit-Distanz zwischen den folgenden Wörtern:

- Datenbank und Datenschrank
- Dateianhang und Karteischrank
- Physiologisch und Psychologisch

Prüfen Sie Ihr Ergebnis, wenn möglich, mithilfe eines SQL-Statements.

5. Ähnlichkeitsbestimmung II

Bestimmen Sie die 4-Gramme der Wörter aus der letzten Aufgabe. Geben Sie die Anzahl übereinstimmender 4-Gramme für jede Wortkombination an. Welches Paar ist am ähnlichsten? Das Ähnlichkeitsmaß beruht auf einem Vergleich der Menge der gemeinsamen 4-Gramme mit den 4-Gramm-Mengen der einzelnen Wörter (Dice-Koeffizient).

6. Ähnlichkeitsbestimmung III

Welchen Soundex haben die Wörter aus Aufgabe 4. Welches Paar ist am ähnlichsten?

7. Differential Snaphshot / Record Linkage

Was wird unter dem Differential-Snapshot-Problem und Record Linkage verstanden? Worin liegt der Unterschied? Die triviale Herangehensweise besteht in der vollständigen Evaluierung sämtlicher Tupel (Jeder-gegen-Jeden).

- Wie kann der Aufwand für die Eliminierung von Duplikaten reduziert werden?
- Welche Vor- und Nachteile haben die einzelnen Techniken?

8. Transformation (SQL)

Überführen Sie mittels SQL-Statements die Relationen *Bierladen1* und *Bierladen2* in die Zielrelation *IntegratedBierladen*.

Bierladen1

PersonalID	Name	Fachrichtung	Abschluss
1	Mark	Verkaeufer	Lehrling
2	Peter	Lagerist	Geselle

Bierladen2

PersonalID	Name	Verkaeufer	Lagerist
1	Mark	Lehrling	NULL
2	Peter	NULL	Geselle

IntegratedBierladen

PersonalID	Name	Geselle	Lehrling