



Aufgabe 4 – Schritt T(ransform)

Um die Vereinsdaten zusammenzuführen, müssen diese nun in der „Staging Area“ zusammengeführt werden. Dieser Schritt **T** ist somit eigentlich der „Kern“ eines ETL Prozesses und soll nun von Ihnen „per Hand“ implementiert werden. Entwerfen / erweitern Sie dazu ein/Ihr Konsolenprogramm in C#, welches das Ergebnis Ihres vorigen Schrittes **E** verwendet (al

so POCO Instanz(en) der drei Datenformate) und als Ergebnis eine vereinigte Form/Darstellung der Daten erstellt, d.h. alle drei Formate in ein einheitliches Datenschema transformiert.

Konkret müssen Sie dabei folgende Probleme lösen¹:

- **Definition eines gemeinsames Zielformat:** Falls nicht in „E“ bereits geschehen oder modelliert, muss ein „Zielformat“ definiert werden, d.h. eine Struktur, in die die zusammengeführten Daten überführt werden (können).
- **Syntaktische Transformation:** Anpassung der Formate/Datentypen (z.B. errechnete Werte und/oder abgeleitete Werte in „neue“ Spalten)
- **Semantische Transformation:** Hierbei werden die Daten auf inhaltliche Aspekte überprüft und wenn nötig modifiziert und angereichert (z.B. unterschiedliche Kodierung von gleichem Inhalt, Duplikate entfernen/vereinigen).

Anforderungen/Kriterien zur Bewertung des Programms

Ihr Programm...

- ...implementiert/definiert ein Zielformat, welches alle Eingabedaten ohne „Verlust“ aufnehmen kann und alle Informationen aller Eingabeformate speichern kann.
- ...setzt die syntaktische Transformation korrekt um, so dass keine Daten(inhalte) verloren gehen und zusätzlich redundante Spalten/Daten vermieden werden.
- ...führt eine umfangreiche semantische Transformation durch, so dass gleiche Informationen als gleiche Daten abgelegt werden (z.B. Vereinigung und Duplikatentfernung)

¹ Siehe auch: https://en.wikipedia.org/wiki/Extract,_transform,_load#Transform

Heterogene Datenquellen zusammenführen - Transform

REU/SDE/PIE

- [optional] ...ermöglicht einem Nutzer interaktiv semantische Transformationen durchzuführen, indem Ihr Programm verschiedene Vorgehensweise vorschlägt (z.B. mögliche Ergebnisse/Schritte für das Zusammenführen von Spalten).
- [optional] ...gibt Statistiken darüber aus, wie „kompakt“ die Daten/Informationen durch die Transformation gespeichert werden konnten, d.h. das Verhältnis der Eingabedaten /-formate in Vergleich zur Ausgabe (also ein Art „Kompressionsfaktor“).
- ...protokolliert den gesamten Prozess in einer sinnvollen und nachvollziehbaren Form (z.B. in Form einer Log-Datei / Konsolenausgaben)
- ...ist vollständig kommentiert, insbesondere Begründungen bei Entscheidungen und/oder Umsetzungen von z.B. einzelnen Teile der Überprüfung der Daten oder der Datenstruktur in der Staging Area.