Aufgabenblatt 5
ETL

- Letzter Abgabetermin: **Montag, 09. Juli 2007**
- Das Aufgabenblatt gilt als bestanden, wenn beide Aufgaben bearbeitet und mindestens 20 der 40 Punkte erreicht werden.
- Abgabe: Per E-Mail an Alexander Albrecht, Fachgebiet Informationssysteme.

Das Ziel dieses Aufgabenblatts ist es, Kundendaten mit Hilfe des IBM Information Servers in einem ETL Prozess zu reinigen und Adressduplikate der Kunden zu entfernen. In dem Beispielszenario will eine Bank ihre Kunden über neue Finanzdienstleistungen informieren. Dabei soll nur ein Angebot pro Haushalt verschickt werden. Die Bank erfasst jedoch Kundeninformationen über die jeweiligen Konten.

Beispiel: Ein Ehepaar hat vier Konten (zwei Girokonten, private Altersvorsorge, Aktienfond). Für diesen Haushalt liegen somit vier Adresseinträge in der Datenbank vor. Die Kundendatenbank kann also Duplikate für Namen und Adressen identischer Haushalte enthalten.

Das Beispielszenario ist als Tutorial in der WebSphere QualityStage Dokumentation zu finden. Während der Übungsveranstaltung wird es eine Einführung des IBM Information Servers geben. Das Standardisieren der Adressdaten (Standardize Stage) als erster Schritt des ETL Prozesses wird demonstriert und die vorgestellten Techniken können parallel an den Übungsrechnern nachvollzogen werden. Eure Aufgabe ist es, in den standardisierten Adressdaten Duplikate zu finden und den „besten" Adressdatensatz für jeden Haushalt bereitzustellen.

Aufgabe 1: Unduplicate Match Stage

Erstelle die *Unduplicate Match Stage* zum Entfernen der Adressduplikate. Verwende dafür die beim Standardisieren generierten Datensätze.

a) Innerhalb der Unduplicate Match Stage werden Datensätze gruppiert, die ähnliche Attributwerte besitzen. Dafür wird eine *Unduplicate Match Specification* verwendet. Erstelle für dieses Beispielszenario mit dem *Match Designer* die *Unduplikate Match Specification*. Das Vorgehen ist im WebSphere QualityStage User Guide, Kapitel 6, zu finden und wird in der Übungsveranstaltung kurz vorgestellt. **10 P**

b) Erstelle die vollständige *Unduplicate Match Stage*. Das Vorgehen ist in Kapitel 5 des WebSphere QualityStage Tutorials beschrieben. Dafür wird die in der vorangegangenen Teilaufgabe erstellte *Unduplicate Match Specification* benötigt. Bei Bedarf wird diese als Musterlösung zur Verfügung gestellt. **10 P**

Aufgabe 2: Survive Stage

In dieser Aufgabe erstellt ihr die *Survive Stage*, die den besten Adressdatensatz für jeden Kunden bereitstellt. Das Vorgehen ist in Kapitel 6 des WebSphere QualityStage Tutorials beschrieben. Als Eingabe dienen die bereinigten Adressdatensätze aus der *Unduplicate Match Stage*. Die bereinigten Datensätze findet ihr auch im Übungsverzeichnis R:\InfoInt07_Naumann\Uebung\0703\. **20 P**