DATAKWALITEIT: DEEL 1

INLEIDING

In dit derde deel van het project willen we het ETL-proces stelselmatig uitbreiden door controles op datakwaliteit in te voeren. Specifiek zullen we hier kijken naar controles op basis van een aantal regels die we willen invoeren.

KWALITEITSPROBLEMEN

In dit deel van het project gaat onze aandacht naar een aantal specifieke problemen met kwaliteit van data in de brondatabanken. Meer bepaald willen we nagaan of (1) gegevens van klanten consistent zijn en (2) gegevens van activiteiten consistent zijn. In het eerste geval zullen we vooral kijken naar adresgegevens en in het tweede geval vooral naar de gegevens in verband met toegankelijkheid voor mensen met een beperking. In beide gevallen is het basisidee om een aantal regels te gaan opstellen die toelaten om consistentie van gegevens te valideren. Je kan hiervoor inspiratie putten uit de verschillende technieken die in de cursus aan bod zijn gekomen.

TECHNISCHE IMPLEMENTATIE

De kwaliteitscontroles moeten worden geïmplementeerd als onderdelen van het ETL-proces. Je kan dit in een aparte stap implementeren. Bij de implementatie is het vooral van belang om je te richten op het kunnen detecteren van problemen. Het is uitdrukkelijk niet de bedoeling om een complete bibliotheek te schrijven met alles erop en eraan. Indien je bijvoorbeeld edit regels gebruikt, is het niet nodig om code te schrijven die een voldoende verzameling genereert (je kan handmatig de voldoende verzameling berekenen). Wel is het nodig om bijvoorbeeld code te hebben die een set van edit regels kan controleren op de gegeven data. Je *kan* in dit deel al nadenken over hoe je de gevonden problemen kan oplossen, maar dat hoeft nog niet.

VERSLAG

In het uiteindelijke verslag van je project, rapporteer je voor dit deel hoe je te werk bent gegaan en wat je bevindingen zijn. Welke regels heb je opgesteld om de consistentie na te gaan? Welke soort regels heb je hier gebruikt? Heb je externe databronnen geraadpleegd? Hoeveel inconsistenties heb je gedetecteerd? Heb je enig inzicht in bepaalde foutmechanismen die optreden? De implementatie van de controles voeg je toe als bijlagen aan het eindverslag.