DATAKWALITEIT: HERSTEL VAN FOUTEN

INLEIDING

In dit vijfde en laatste deel van het project willen we de resultaten van deel 3 en deel 4 gaan benutten om de kwaliteit van gegevens in de verschillende tabellen in het warehouse te gaan verbeteren. Dit doen we door hersteltechnieken toe te passen.

HERSTEL VAN INCONSISTENTIES

In deel 3 kreeg je de opdracht om in de data van bezoekers en activiteiten te zoeken naar mogelijke inconsistenties. Met dit systeem kan je rijen opsporen waar (mogelijks) fouten aanwezig zijn. In dit deel is het de bedoeling om het herstel van deze fouten door te voeren. Hiervoor kan je gebruik maken van de technieken die in de lessen aan bod zijn gekomen. Gebruik je edit regels, dan kan je een minimale set cover zoeken en voor de gekozen attributen nieuwe waarden bepalen. Gebruik je afhankelijkheden, dan kan je gebruik maken van het Chase algoritme. Gebruik je andere technieken, zoals bijvoorbeeld referentiedata, dan kan je meestal de gekende technieken omvormen zodat ze toepasbaar worden.

Implementeer je methode voor het herstel van fouten en voeg deze stap toe aan je ETL-proces. De bedoeling is dat je de correcties effectief doorvoert in je warehouse.

VERWERKEN VAN DUBBELS

In deel 4 kreeg je de opdracht om te zoeken naar dubbele data binnen de bezoekersdata. In het huidige deel gaan we deze dubbele data verwerken. Dit gaan we doen door dubbele data samen te voegen. Dat betekent dat we verschillende bezoekers die we als dubbel hebben geïdentificeerd gaan samenbrengen in één rij. Maak gebruik van de technieken uit de syllabus om dit te doen.

Ook in deze stap is het de bedoeling je code uit te breiden zodat dubbele voorkomens van bezoekers worden verwerkt en als één instantie worden opgeslagen in het warehouse.

VERSLAG

In het uiteindelijke verslag van je project rapporteer je voor dit deel hoe je te werk bent gegaan voor beide stappen. Geef uitleg over hoe je precies correcties hebt gekozen. Voor het verwerken van dubbele data: bespreek je strategie om het samengevoegde tuple samen te stellen. Geef ook een korte uitleg over de (on)zekerheid dat je een juiste beslissing hebt genomen. Voeg je code, zoals steeds, toe als bijlage aan het eindverslag.