Data Kwaliteitsrapport MHL

Een analyse van de data-integriteit van de MHL database

Introductie

De integriteit van de data van MHL kan door intensief gebruik en applicatiewijzigingen niet meer gewaarborgd worden. In dit rapport wordt de data-integriteit beoordeeld op vier maatstaven: *Entity Integrity, Referential Integrity, Domain Integrity* en *Constraint Integrity*. Na iedere bevinding volgen de stappen om de kwaliteit van de data weer op acceptabel niveau te krijgen.

Bevindingen & Adviezen

Entity Integrity

Deze vorm van integriteit betreft de conformiteit naar de derde normaalvorm van Codd [1]. Deze is over het algemeen in orde. Een belangrijk element van de normaalvormen van Codd is echter dat data die afgeleid kan worden geen eigen kolom krijgt. In dit licht is het aan te raden om de `property_DEF` kolom in `mhl_detaildefs` te verwijderen, aangezien deze informatie gesloten zit in `display_name`. In theorie kunnen ook de postcodes in `mhl_suppliers` worden achterhaald aan de hand van de straat en het huisnummer. Dit heeft alleen veel voeten in de aarde omdat een API abonnement moet worden afgesloten en dit moet worden geïmplementeerd. Een ander aspect van de normaalvormen van Codd is dat een cel maar een waarde mag hebben. De `properties` in `mhl_detaildefs` zijn arrays. Daarnaast is hun gebruik niet goed opgenomen, aangezien `propertytype_ID` soms wel is aangegeven maar niet in de array met `properties` staat, en andersom. Met behulp van een extra tabel zijn deze groepen gemakkelijk te ontwerpen in de database. Tot slot heeft de tabel `mhl_hitcount` geen Primary Key (PK). De huidige PK maakt gebruik van drie waarden. Het loont de moeite om hier ook een `id` kolom toe te voegen met een Auto-Increment functie (AI).

Referential Integrity

Deze vorm van integriteit slaat op de onderlinge relaties tussen tabellen. De Referential Integrity is voor de MHL database **zwaar ondermaats.** Er zijn vrijwel geen enkele Foreign Keys (FK) opgenomen terwijl die er wel moeten zijn. In het Entity Relation Diagram zijn alle nieuwe relaties weergegeven en staat bij iedere kolom of dit een FK is. Een script zal later geleverd worden om deze relaties en *constraints* toe te voegen aan de database. Ook valt op dat de `mhl_brands` tabel geen relatie lijkt te hebben met de rest van de data, waar dit wel zou kunnen qua semantiek.

Domain Integrity

Domain Integrity betreft de correctheid van de datatypen en het uitsluiten van *outliers.* Het gros van de datatypen in de MHL database zijn correct, echter vallen een aantal

onregelmatigheden op. Zo zijn `vlevel` in `mhl_detaildefs` en `ulevel` in `mhl_membertypes` een `int(11)` maar hebben beide kolommen alleen waarde 1 of 2. Om opslagruimte te besparen kan dit worden teruggebracht naar een `tinyint`. Ook komen de datatypen van `pc6` in `pc_lat_long` (`char(6)`) niet overeen met de datatypen voor de postcodes in `mhl_suppliers` (`varchar(7)`). Uniformering is dus geboden, zeker omdat ze een relatie met elkaar hebben. Deze relatie is echter niet uitgedrukt met een primary key en foreign key. Tot slot heeft de `content` kolom in `yn_properties` alleen de waarde “Y”. Deze kolom is dus momenteel overbodig.

Constraint Integrity

Deze vorm van integriteit betreft de eisen die aan de data input worden gesteld. De huidige eisen laten bij veel tabellen nog te wensen over. Zo worden op veel plekken lege strings gebruikt als `NULL` waarden. Het is beter om `NULL` waarden in te voegen als een veld optioneel is, zoals bij het `contacttype`, `email` en `tel` in `mhl_contacts`. Hetzelfde geldt voor de `straat`, `huisnummer`, `p_address` en `p_postcode` kolommen in `mhl_suppliers`. Ook missen de `countrycodes` van België, Duitsland, Australië, Denemarken, Oostenrijk en China. Op een aantal plekken komen ook lege strings voor waar onduidelijk is of deze optioneel zijn of niet. Deze zijn: `content` in `mhl_properties`, `display` in `mhl_propertytypes` en `content` in `yn_properties`.

Conclusie

De MHL database bevindt zich in slechte staat. Op ieder gebied van data-integriteit zijn fouten te bespeuren met als grootste defect het ontbreken van FKs. Het bijgeleverde script kan hier orde op zaken stellen. Het is ten harte aan te bevelen om de overige gesuggereerde wijzigingen in dit document over te nemen.

Bronnen

[1] Codd, E. F. "Further Normalization of the Data Base Relational Model". (Presented at Courant Computer Science Symposia Series 6, "Data Base Systems", New York City, May 24–25, 1971.) IBM Research Report RJ909 (August 31, 1971). Republished in Randall J. Rustin (ed.), *Data Base Systems: Courant Computer Science Symposia Series 6*. Prentice-Hall, 1972.