Referentiële data onderscheidende datalaag binnen informatie-architectuur

Introductie in de wereld van referentiële data

Malcolm Chisholm

De term Reference Data, of referentiële data, wordt door verschillende mensen verschillend geïnterpreteerd. Hierdoor ontstaat spraakverwarring als IT-specialisten er met elkaar over spreken. In dit artikel wordt een precieze definitie van referentiële data gepresenteerd, waarna gekeken wordt naar de diverse soorten die binnen de omschrijving vallen.

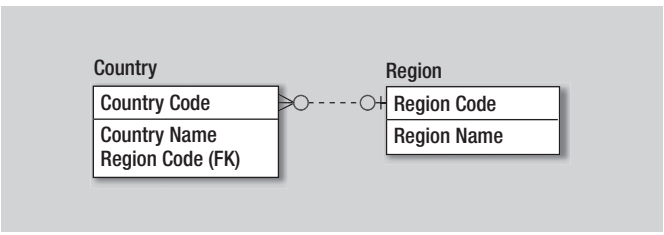
Omdat een toenemend aantal organisaties zich buigt over beter beheer van hun referentiële data, is het belangrijk om de term goed te omschrijven en te begrijpen. Een beter begrip draagt niet alleen bij aan een betere communicatie, maar maakt het ook gemakkelijker in te zien welke specifieke beheertaken moeten worden toegepast.

Definitie

Omdat er geen bepaalde eenduidige definitie bestaat van referentiële data, heb ik in mijn boek 'Managing Reference Data in Enterprise Databases' (Morgan Kaufmann, 2000) een poging gedaan:

Referentiële data betreffen alle soorten data die uitsluitend gebruikt worden om andere data in de database in categoriseren, of die uitsluitend data in de database relateren aan informatie buiten de ondernemingsgrenzen.

Onder deze betekenis van referentiële data vallen dan dus ook bijvoorbeeld 'lookup data' of 'domain values'. Referentiële data die onder deze definitie vallen worden in tabellen geplaatst die meestal twee kolommen hebben, hoewel er soms ook meer dan twee gebruikt worden. Deze tabellen hebben gewoonlijk maar een paar rijen, vaak minder dan tien, maar bijna altijd minder dan 200. Afbeelding 1 toont een klein gedeelte van een data-model voor het ontwerp van twee referentiële data-tabellen.



Afbeelding 1: Typisch voorbeeld van referentiële data-tabellen.

De entiteit *Country* heeft een *Country Code* sleutelattribuut en een *Country Name* non-sleutelattribuut. Referentiële data-tabellen hebben bijna altijd een 'code'-kolom als primaire sleutel, hoewel er soms nog toegevoegde primaire sleutelkolommen kunnen zijn. Een code is kort, maar de standaardmanier om een voorbeeld van een referentiële data-entiteit te identificeren; voor de code wordt bijna altijd de voorkeur gegeven aan een acroniem boven de betekenisloze surrogaat-sleutel zoals een volgnummer.

In afbeelding 1 behoort een *Country* uitsluitend tot één *Region*. De entiteit *Region* heeft ook een code als primaire sleutel en een 'beschrijvende' kolom (Region Code) als niet-sleutel. Vaak denken datamodelleerders dat referentiële data-tabellen geen relaties hebben en alleen bestaan uit een 'code'-kolom (de primaire sleutel) en een enkele 'beschrijvende' niet-sleutelkolom. In de praktijk is het echter gebruikelijk dat er relaties zijn, zoals die bijvoorbeeld getoond worden in afbeelding 1. Binnen deze relaties heeft een referentiële data-entiteit één of meer referentiële data-entiteiten als 'Ouder'; de relaties zijn bijna altijd niet-identificerend. Belangrijk te vermelden is, dat elke 'Ouder' van een referentiële data-tabel zelf ook altijd een referentiële data-tabel is. Binnen een datamodel hebben referentiële data-entiteiten altijd relaties met andere 'Kind'-entiteiten. Terwijl deze 'Kind'-entiteiten soms andere referentiële data-entiteiten zijn, zijn ze meestal entiteiten die geen referentiële data zijn. Een Valuta-entiteit is bijvoorbeeld een typische referentiële data-entiteit, die vele relaties kan hebben met andere entiteiten die attributen met financiële gegevens bevatten. Voor elk financieel gegeven is het noodzakelijk om de bijbehorende valuta te achterhalen. Dit is één van de redenen waarom datamodellen vaak onleesbaar zijn – ze hebben te veel rijen die relaties voorstellen die voortkomen uit 'Ouder'-referentiële data-entiteiten.

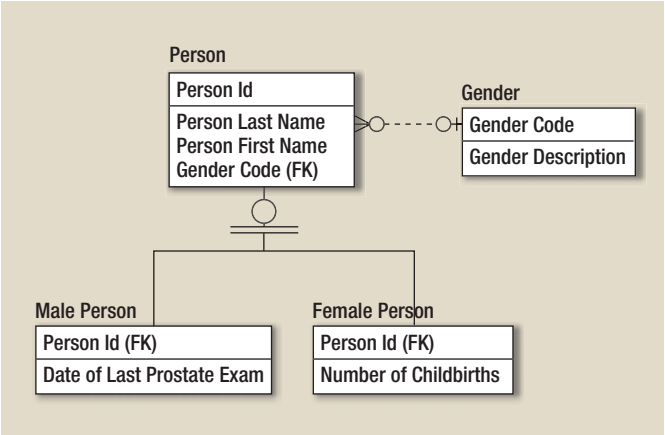
In de praktijk zorgen veel datamodelleerders er altijd voor om

'subject areas' te maken die de referentiële data-entiteiten en hun relaties verbergen, zodat ze zich kunnen concentreren op de meer 'belangrijke' entiteiten.

Verscheidenheid

Terwijl het mogelijk is referentiële data te definiëren, en te begrijpen dat referentiële data-tabellen dezelfde algemene structuur hebben, is het soort data dat in de referentiële data-tabellen is vastgelegd, behoorlijk verschillend. Er zijn drie hoofdonderscheiden te maken:

- *Externe referentiële data*. Dat zijn data die de organisatie niet zelf beheert en die dus uit bronnen komen buiten de grenzen van de organisatie. *Country* en *Currency* zijn daarvan goede voorbeelden. Geen enkele onderneming 'beheert' landen en valuta's; er zijn internationale standaards voor beide referentiële data-tabellen (bijvoorbeeld ISO). Het is duidelijk dat deze tabellen geen echte eigenaar binnen een organisatie hebben omdat de data uit externe bron komen.
- *Structurele referentiële data*. Sommige referentiële data worden op bepaalde manier gebruikt in een door de database afgedwongen architectuur. Bijvoorbeeld, als een supertype verscheidene subtypes heeft, moet er een referentiële data-tabel zijn die één record van elk subtype bevat. Deze wordt de *Category Discriminator* genoemd. Afbeelding 2 toont een fragment van een model van een medische database, waarin de entiteit *Gender* de *Category Discriminator* is voor de entiteit *Person*. De tabel *Gender* zal twee records bevatten: één voor *Male Person* en een ander voor *Female Person*. Andere referentiële data-tabellen die bepalend zijn voor de architectuur, zijn bijvoorbeeld die statuscodes bevatten voor de levenscyclus van een transactie. Voor een transactie die betrokken is op de verkoop van een product, kan er een referentiële data-tabel zijn die telkens één record bevat uit *Order Placed*, *Payment Received, Product Shipped* en *Product Received*. Dit type referentiële data wordt vervaardigd door IT-personeel dat betrokken is bij ontwerp. Deze medewerkers kunnen de echte autoriteit zijn voor dergelijke tabellen, zelfs als andere personen binnen de organisatie formeel eigenaar van de gegevens zijn.
- *Taxonomiën*. Deze representeren de derde hoofdgroep van referentiële data. Taxonomiën classificeren andere soorten data in de database. Vandaag de dag lijkt het woord 'taxonomie' te zijn voorbehouden aan de wijze van indexering van ongestructureerde data zoals documenten. Classificatieschema's zoals taxonomiën komen echter zeer veel voor in databases. De entiteit *Regio* in afbeelding 1 is bijvoorbeeld een taxonomie. Er bestaat geen objectieve set regio's: verschillende organisaties hanteren verschillende regio's. In de praktijk hebben zelfs verschillende bedrijfsonderdelen hun eigen vastgestelde regio's, wat problemen geeft als de data geconsolideerd moeten worden op ondernemingsniveau. Taxonomiën worden meestal gebruikt als rapportage-dimensie; er is geen bovengrens aan het aantal dat in de database kan worden ingevoerd. Vaak leiden nieuwe rages en modes tot de introductie van aanvullende taxonomiën.



Afbeelding 2: Voorbeeld van een Category Discriminator.

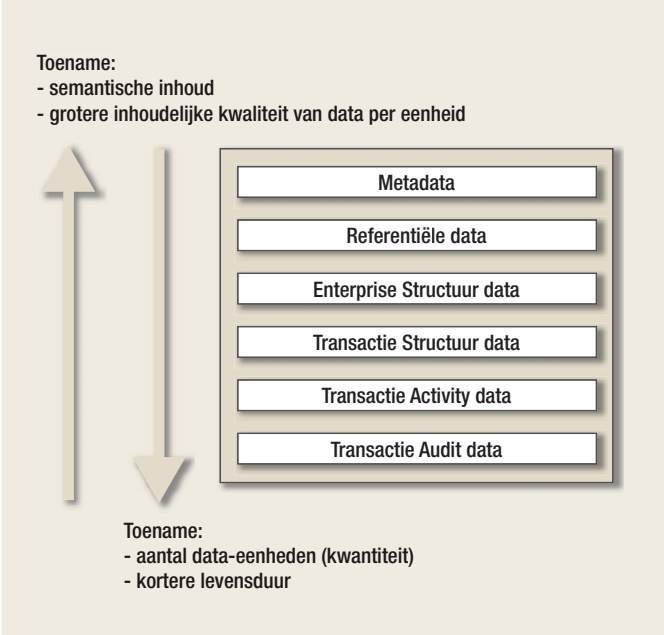
Externe referentiële data, structurele referentiële data en taxonomiën zijn de drie hoofdgroepen van referentiële data. Elk heeft verschillende eigenschappen en stelt eigen eisen aan het beheer. Hoe het ook zij, ze vallen binnen de eerdergenoemde algemene definitie van referentiële data en worden ingevoerd met de kenmerkende, eenvoudige 'code & description'-structuur van referentiële data-tabellen. We gaan nu eens kijken waar we referentiële data kunnen toepassen in de ondernemingsbrede informatie-architectuur.

Datalagen

Een van de problemen van het begrijpen van referentiële data lijkt te stammen uit de datamodellen. Binnen een datamodel ziet elke entiteit er hetzelfde uit. Dit wekt de indruk dat alle entiteiten binnen het datamodel vrijwel gelijk zijn in termen van beheer-activiteiten. Een datamodel maakt niet echt onderscheid tussen referentiële data-entiteiten en andere entiteiten en suggereert niets over de bijzondere rollen die referentiële data-entiteiten spelen. Toch is het mogelijk om de data in te delen in zes categoriën, zes lagen die verschillende rollen spelen en hun eigen specifieke beheer-eisen stellen. Deze zes lagen zijn te zien in afbeelding 3.

Elk van de lagen wordt hieronder verder uitgelegd.

- Metadata. Dit zijn data die de organisatiebrede informatie-architectuur beschrijven, zoals definities van tabellen en kolommen in de systeem-catalogus van een database. Metadata worden niet altijd getoond in de database/tabellen waar de gebruikers mee werken, maar soms wel.
- Referentiële data. Zoals eerder beschreven, zijn referentiële data alle typen data die worden gebruikt om andere data in de database te categoriseren, of alleen om data in een database te koppelen aan informatie buiten de grenzen van de onderneming.
- Enterprise Structuur data. Dit zijn data die de structuur van de onderneming beschrijven, bijvoorbeeld organisatiestructuur of boekhouding. Deze informatie wordt gebruikt om bedrijfs-



Afbeelding 3: De zes datalagen.

activiteiten per verantwoordelijkheidsgebied te volgen.

- Transactie Structuur data. Deze (ook wel 'master data' genoemd) beschrijven de partijen betrokken bij de transacties van de onderneming. *Product* en *Customer* zijn twee gebruikelijke entiteiten in deze categorie. Als zich een transactie voordoet, zoals een verkoop of een activiteit op het gebied van human resources, moeten de data in deze tabellen aanwezig zijn voor de transactie kan worden afgesloten. Vaak worden transacties gereguleerd door wetten, voorschriften of contracten die het wie en wat bepalen van de partijen in een transactie.
- Transactie Activity data. Dit is de traditionele focus van IT. Het betreft de data die de transacties vormen, die door de operationele systemen van de onderneming worden behandeld zoals verkopen en activiteiten op het gebied van human resources.
- Transactie Audit data. Een individuele transactie kan verschillende fasen doorlopen. In elke fase kan de status veranderen. Audit informatie volgt deze statusveranderingen. Web logs en database logs volgen dit soort data ook.

Deze zes datalagen staan in afbeelding 3 op volgorde van belangrijkheid van individuele datawaarden. Als bijvoorbeeld een enkel metadata-item, zoals het datatype van een kolom, zou worden gewijzigd, heeft dat zeer vergaande gevolgen. Daartegenover staat dat elke dag grote hoeveelheden transactie audit data gewist worden uit de log files. De hogere niveaus's in afbeelding 3 hebben meer semantische inhoud in termen van individuele datawaarden en dat is zeker het geval bij referentiële data.

Een voorbeeld van de grote semantische inhoud van referentiële data-waarden, is dat deze waarden vaak gebruikt worden in business rules. Het blijkt zelfs dat als een business rule expliciete datawaarden bevat, deze altijd uit referentiële data-tabellen komen. Ergo, als de waarde van referentiële data worden veranderd of gewist, kan dit de uitvoering van deze business rules stopzetten en onvoorspelbare effecten hebben op het onderliggende systeem.
Omdat er op dit moment geen eenvoudige manier is om referentiële data te relateren aan business rules die zijn ingevoerd in bijvoorbeeld de programma broncode, zal dit probleem nog wel enige tijd aanhouden. Het verklaart waarom IT-medewerkers vaak bang zijn om de waarden van referentiële data te wijzigen of te wissen. In veel organisaties mag niet worden gewijzigd in de primaire sleutelwaarden in de referentiële data-tabellen, hoewel het toevoegen van nieuwe records over het algemeen geen problemen oplevert.

Afbeelding 3 laat ook de manier zien waarop de term 'referentiële data' breder gebruikt kan worden. Sommige IT-professionals beschouwen ook Enterprise Structure Data en Transaction Structure Data als onderdeel van referentiële data. In de praktijk luidt de definitie van referentiële data dan "data die wordt geconsumeerd door een systeem maar niet gemaakt en beheerd wordt door het systeem." Zo'n definitie is veel te breed en identificeert niet echt data-categoriën met unieke eigenschappen en beheereisen. Daardoor kunnen projecten die dit soort 'referentiële data' proberen te managen geen gewone beheertechnieken vinden die succesvol kunnen worden toegepast. Het is beter om de structuur uit afbeelding 3 te gebruiken en beheertechnieken te gebruiken die aansluiten bij de unieke eigenschappen van elke laag.

Conclusies

Referentiële data zijn een onderscheidende datalaag binnen informatie-architectuur van de onderneming, met eigen specifieke eigenschappen en beheereisen. Referentiële data moeten niet worden verward met andere datatypen. Referentiële data kunnen worden onderverdeeld in minstens drie hoofdgroepen en elk daarvan heeft eigen unieke eigenschappen en beheereisen.
Omdat steeds meer organisaties proberen hun referentiële data te beheren, moeten zij zich bewust zijn van de betekenis als ze de term 'referentiële data' gebruiken en zich te focussen op een aanpak die zich richt op deze dataklasse.

Malcolm Chisholm is onafhankelijk consultant en auteur. Hij is tevens spreker op het congres Database Systems.

Mastering Reference Data

Malcolm Chisholm geeft op het Database Systems congres een tutorial op 25 maart.
Deze brede tutorial behandelt naast de theorie ook praktische aspecten. Zie voor meer informatie pagina 25 en 27.