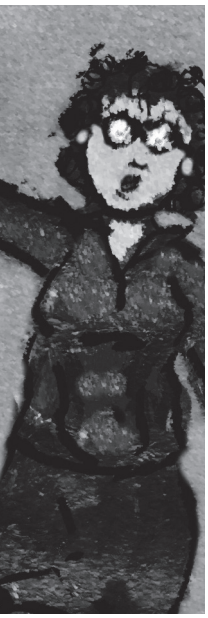


Het probleem van meervoudige identificatie belicht

Eindelijk grote belangstelling voor Master Data

Malcolm Chisholm

De laatste tijd neemt de belangstelling voor Master Data flink toe, na vele jaren waarin er nauwelijks of geen aandacht was voor het onderwerp. Uit door mij gedaan onderzoek blijkt dat de drijvende kracht hierachter meer de gebruikers zijn dan de IT'ers.

De precieze reden voor de plotselinge vraag naar Master Data Management (MDM) is niet duidelijk, maar men blijkt zich in toenemende mate te realiseren dat ondernemingen Master Data volledig moeten beheersen, als ze data effectief willen kunnen delen in applicaties, variërend van datawarehouses tot 'rechtoe-rechtaan' processing. Het is ook duidelijk geworden dat MDM nauw verbonden is met datakwaliteitsproblemen in het algemeen. Het is mogelijk dat die problemen de gebruikers pas duidelijk zijn geworden na de implementatie van een datawarehouse en/of datamarts, of nadat voormalige stand-alone systemen werden gekoppeld aan procestransacties.

Wat de onderliggende redenen ook mogen zijn, er is op dit moment een enorme vraag naar MDM. Een groot probleem is echter, dat veel IT'ers zich niet realiseren dat de klasse Master Data verschilt van andere dataklassen en dat de unieke eigenschappen van MDM speciale oplossingen verlangen.
Bij verschillende gelegenheden werd mij door IT'ers gezegd dat MDM niet afwijkt van datamanagement in het algemeen, dat het niet veel meer is dan een voorbijgaande modegril. Daar ben ik het mee oneens. Master Data hebben unieke eigenschappen en gedragingen, die hun managementbehoefte dicteren. Deze eigenschappen en gedragingen worden niet aangetroffen bij andere dataklassen en men kan niet volstaan met een globale aanpak of een *one-size-fits-all* benadering van MDM. Projecten waarin getracht wordt MDM op deze manier te implementeren, zijn gedoemd te mislukken.

Een van de fundamentele problemen met Master Data is: wat identificeren ze. Dat klinkt misschien raar voor iemand die niet goed bekend is met MDM, maar dit bezorgt de meeste hoofdpijn en kan gemakkelijk de manier beperken waarop een bedrijf Master Data als corporate asset kan aanwenden.
Voordat we dit probleem nader gaan beschouwen, is het noodzakelijk om te definiëren wat Master Data zijn. Afbeelding 1 geeft een precieze definitie van Master Data, samen met definities voor Reference Data (referentiële/referentie data, ook bekend als 'lookup data' en 'domain values') en Enterprise Structure Data. In de praktijk wordt meestal alles uit afbeelding 1 'Master Data' genoemd. In dit artikel zal ik de exacte definitie gebruiken van Master Data, zijnde "the data that represents the parties to the transactions that record the operations of an enterprise."

Dataklasse	Definitie	Voorbeeld
Master Data	Data die de directe deelnemers in een transactie vertegenwoordigen, en die aanwezig moeten zijn voordat een transactie plaatsvindt.	Klant, product
Reference Data	Elk type data dat alleen wordt gebruikt voor de categorisatie van andere data uit een database, of alleen om data uit een database te relateren aan informatie buiten de grenzen van de onderneming.	Land, valuta, productlijn, klanttype
Enterprise Structure Data	Data die de structuur van de onderneming zodanig beschrijven, dat de bedrijfsactiviteiten per bedrijfsverantwoordelijkheid kunnen worden gerapporteerd.	Organisatorische structuur, overzicht boekhouding

Afbeelding 1: Definities van Master Data en gerelateerde dataklassen.

Algemene voorbeelden daarvan zijn Klant en Product. Data daaromtrent moeten worden vastgelegd, voordat een transactie waarbij een Product aan een Klant wordt verkocht, kan worden uitgevoerd.

IT heeft zich traditioneel gefocust op de transactie en de bijbehorende data, in plaats van op de data op hogere abstractieniveaus, zoals Master Data, Reference Data en Metadata. Dat kan prima in silo-systemen waarbij de data nooit worden uitgewisseld, maar geeft problemen als men probeert enige vorm van data-sharing of integratie toe te passen. Het is een nog groter probleem als niemand weet hoe een bepaald onderdeel van Master Data heet.

Wat is een Ding?

We nemen een product als voorbeeld. Je zou denken dat elke onderneming elk product op een unieke manier zou identificeren. Dit is echter zelden het geval, om uiteenlopende redenen.
Een belangrijke factor is dat een product altijd een levenscyclus doorloopt en in elke levensfase verschillende namen heeft. Het kan een idee zijn, dan een prototype. Misschien is de volgende fase de productie van het product. Op een zeker moment kan de productie worden gestaakt, terwijl de bij de garantie horende serviceverlening nog doorloopt. Uiteindelijk kan ook die stoppen, maar kan de onderneming nog steeds verplichtingen voor het product hebben. Gedurende al deze stappen in de levenscyclus kan het product een andere identificator hebben. Dit soort dingen drijft data-administrators tot wanhoop. Hun standpunt is dat een product moet worden voorgesteld door één unieke identificator, één enkele datawaarde, en dat iemand alleen maar even hoeft te bepalen welke dat zal zijn.

Om diverse redenen zal dat nooit gebeuren. Vaak wordt een product geïdentificeerd door een intelligente sleutel (key) die alleen maar mee hoeft te wijzigen met de levenscyclus. Een 'intelligent key' is een sleutel waarvan de datawaarden zodanig worden voorgesteld, dat een mens kan leren hoe die geïnterpreteerd moeten worden. Het negen-cijferige Social Security Number in de VS bijvoorbeeld, heeft als format 123-45-6789, waarbij 123 de geografische aanduiding is, 45 staat voor een chronologische volgorde die ruwweg overeenkomt met de tijd, en 6789 is een volgnummer. Intelligente sleutels bevatten altijd meerdere stukjes informatie. Dit is in strijd met de basisregels voor datamodellering, die immers stellen dat één eenheid data slechts één enkel stuk informatie mag bevatten.

Intelligente sleutels kunnen grenzen hebben, waardoor ze aangepast moeten worden. Ik werkte voor een organisatie die projecten ten uitvoer bracht, en het jaartal was onderdeel van het project-identificatienummer. Soms verstreken er zoveel jaren voordat een project in productie werd genomen, dat het projectnummer moest worden herzien en aangepast aan het jaar waarin het project werkelijk in productie ging. Gezien vanuit marketing- en PR-standpunt is daar niets mis mee, maar we konden nooit matchen tussen de nog niet en de al wel in productie genomen projecten, omdat de projectnummers waren veranderd.

Surrogaatsleutels zijn een vloek

IT-medewerkers hebben de neiging de gebruikers te beschuldigen van onoorbaar gedrag, zoals het veranderen van identificatienummers, maar vaak dragen ze zelf hun steentje bij aan de verwarring. Surrogaatsleutels lenen zich daar bij uitstek voor. Dit soort sleutels zijn niet meer dan willekeurige cijfers of volgnummers, die geen enkel aspect beschrijven van de Master Data waarmee ze verband houden. De meer technische IT'ers, zoals programmeurs, zijn er gek op om éénkoloms identificatienummers te gebruiken als primary key's voor Master Data. Ze denken dat surrogaatsleutels het makkelijker maken om de uniekheid te controleren; bovendien worden zo problemen met multi-kolom sleutels vermeden. In theorie betekent deze oplossing ook dat er geen enkele noodzaak is om het identificatienummer aan te passen aan de voortgang in de product life cycles.

> Er kunnen vanuit de business valide redenen zijn om meerdere identifiers te gebruiken

In werkelijkheid identificeren surrogaatsleutels records in de database-tabel, en stellen geen bedrijfsentiteiten voor zoals individuele producten. Ze hebben geen betekenis voor de business. Vandaar dat surrogaatsleutels waardeloos worden als informatie buiten de grenzen treedt van een bepaalde applicatie, en moeten we vertrouwen op andere kolommen in onze producttabel om te bepalen of een product in applicatie X hetzelfde is als een product in applicatie Y.

Er is nog een probleem met surrogaatsleutels: ze werken niet echt goed voor mensen. Het zijn prachtige dingen voor computers en databases, maar gebruikers hebben behoefte aan identifiers met begrijpelijke informatie. Intelligente sleutels doen dat. Ze kunnen een product uniek identificeren, maar ze kunnen ook aanvullende betekenis hebben als iemand ze buiten de computer gebruikt. Voor gebruikers is het werken met surrogaatsleutels veel moeilijker, omdat deze key's er uitzien als willekeurige getallen. Zo is op Amerikaanse rijbewijzen de geboortedatum van de eigenaar verwerkt in het ID-nummer van het rijbewijs, wat een nuttige echtheids-check kan zijn als het gebruikt wordt als identificatiebewijs voor bijvoorbeeld een lening. Dit zou niet mogelijk zijn als het ID-nummer alleen maar een willekeurige reeks cijfers zou zijn. Als er in een computersysteem geen intelligente sleutels beschikbaar zijn, komen daar vaak problemen van. Onlangs sprak ik met de medewerkers van een grote huizen-

bouwer, die problemen hadden met het traceren van verzoeken van klanten om aanpassingen. Die codes die de aanpassingen – zoals stenen vloeren, kranen, verfkleuren – moesten identificeren, waren surrogaatsleutels, willekeurige getallen. Het gevolg was dat er bij het overschrijven vaak fouten werden gemaakt, omdat de gemiddelde verkoper niet aan de code kon zien of hij te maken had met plavuizen, sanitair, verf of wat dan ook. Als de code intelligent was geweest, hadden de verkopers dat wel kunnen controleren en zouden er veel minder datakwaliteitsproblemen geweest zijn. Dit soort zaken komt zelden ter sprake tijdens technisch overleg over sleutels.

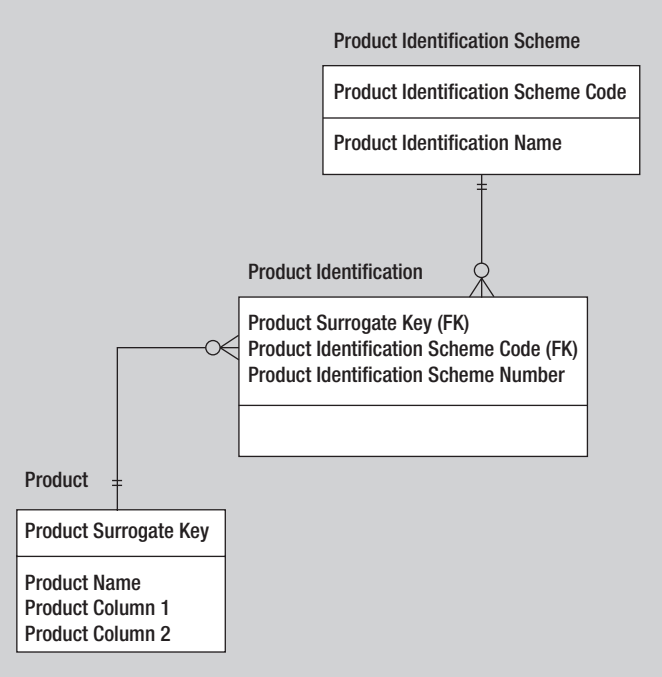
Concurrerende standaards

Het gebruik van standaards is nog een andere manier om producten te identificeren. In de VS dragen aandelen, obligaties en andere financiële instrumenten een identificatiecode die CUSIP heet. CUSIP staat voor Committee on Uniform Securities Identification Procedures; deze standaard is eigendom van de Amerikaanse bankiersvereniging en wordt toegepast door Standard and Poor's. De mastertabellen van de producten van de effectenmakelaars worden gevuld met informatie over de waardepapieren, die geïdentificeerd kunnen worden aan de hand van hun CUSIP-nummers. Helaas is de CUSIP-standaard niet de enige. ISIN (International Securities Identification Numbering System) en SEDOL (Stock Exchange Daily Official List) zijn andere standaards, die zowel worden toegepast op effecten zonder CUSIP-code als op waardepapieren mèt. Dit veroorzaakt een enorm probleem voor organisaties die een breed scala aan beleggingspapieren aanbieden, en zaken doen met klanten die allemaal hun eigen manier van identificatie hanteren. Opnieuw lijken de pogingen om tot één identificatie per product te komen, te worden gedwarsboomd.

> In werkelijkheid identificeren surrogaatsleutels records in de database-tabel

Intuïtief wil je dan je pogingen verdubbelen om de acceptatiegraad van standaards zo hoog mogelijk te krijgen. Maar ook dat is geen garantie voor uniformiteit bij de identificatie van Master Data.

Er kunnen zelfs vanuit de business valide redenen zijn om meerdere identifiers te gebruiken. Winkeliers verstrekken vaak schriftelijke prijsopgaven. Als natuurlijke reactie daarop gaat de klant met de prijsopgave naar andere winkels in de hoop het product goedkoper te kunnen krijgen. Om dat te voorkomen veranderen winkeliers vaak de identifier van de leverancier in een interne identifier. Deze staat met een vage beschrijving van het



Afbeelding 2: De relaties tussen meervoudige product identifiers.

product op de prijsopgave. Slecht nieuws voor de consument die dacht even makkelijk de prijzen te kunnen vergelijken, omdat andere winkeliers nooit kunnen achterhalen op welk product de prijsopgave precies betrekking heeft.

Confrontatie met meervoudige identifiers

Het is een gegeven feit dat er altijd meerdere manieren zullen zijn om een product te identificeren. Het is beter dat database-ontwerpers, DBA's en informatie-architecten de realiteit onder ogen zien, dan er tegen te vechten in een schier oneindige strijd over de puurheid van de identificering. Bovendien is het geen heksenwerk om ontwerpen te maken die kunnen omgaan met meervoudige identifiers voor Master Data. Afbeelding 2 toont een klein stukje database-ontwerp dat dit probeert te doen.

In afbeelding 2 heeft de producttabel een surrogaatsleutel als identifier. De tabel 'Product Identification Scheme' heeft één entry voor elk verschillend schema of standaard om het product te identificeren. In het eerder genoemde voorbeeld over de waardepapieren, zou er een record zijn voor CUSIP, een voor SEDOL en een voor ISIN. Bij een Product Development Life cycle zou er een record zijn voor een laboratorium-ID als het product in de conceptuele fase is, een ander record als het een prototype is, een voor als het in productie is, nog een ander voor als het product niet langer in productie is maar nog wel een garantietermijn heeft, en een record als het product verouderd is. De tabel 'Product Identification' kan meerdere records per product hebben. Hier wordt de surrogaatsleutel gebruikt om records in de Product-tabel te identificeren, in verband gebracht met de verschillende 'Product Identification Schemes' die door de business worden gebruikt. Elk 'Product Identification Scheme' moet een 'Product

Identification Scheme'-nummer hebben, dat het product in kwestie op een unieke manier identificeert.
Een leuk aspect van het ontwerp in afbeelding 2 is, dat er geen record in de tabel 'Product Identification' hoeft te zijn als er geen in de business is. Als een obligatie wel een CUSIP- maar geen ISIN-code heeft, dan is er geen record in de 'Product Identification'-tabel voor ISIN.
Als een Product wel een Laboratorium-ID heeft maar geen Prototype-ID, dan is er geen record in de 'Product Identification'-tabel voor de prototype-fase in de product life cycle. Afbeelding 2 is natuurlijk puur illustratief bedoeld, niet als bruikbaar model. U zult uw eigen ontwerp moeten uitdokteren om binnen uw onderneming met meervoudige aanduidingen van product-identificatie, of andere Master Data, te kunnen werken.

Conclusie

Net als de menselijke natuur, is meervoudige identificatie van Master Data niet perfect, we kunnen dit echter niet veranderen. Daarom is de vraag hoe we ermee kunnen leren leven. De eerste stap is om te erkennen dat meervoudige-identificatieschema's niet alleen maar bestaan om het leven van de IT-medewerkers zuur te maken. Er kunnen binnen de business allerlei redenen zijn waarom meervoudige identificatie er is. Proberen dat ongedaan te maken, kan uitlopen op een lange en moeilijke strijd, en zelfs als er uiteindelijk een overwinning uit de bus komt, kan dit nog meer problemen veroorzaken voor de organisatie. De realiteit is dat meervoudige identificatie voor Master Data bestaat, en dat ermee zal moeten worden gewerkt. Er zijn database-ontwerpen die hieraan kunnen voldoen. Er zitten echter veel meer facetten aan Master Data dan de problematiek van meer-voudige identificatie, en hiermee moet ook worden afgerekend, als u overall Master Data Management wilt bereiken. Het goede nieuws is dat de enorm toegenomen belangstelling voor Master Data Management de aandacht brengt naar het volledige spectrum van aspecten en we kunnen eindelijk echte vooruitgang op dit gebied tegemoet zien na vele jaren van relatieve stagnatie.

> Er zitten veel meer facetten aan Master Data dan de problematiek van meervoudige identificatie

Noot

De oorspronkelijke Engelstalige tekst van dit artikel kunt u vinden op www.dbm.nl in het hoofdmenu onder Specials/Extra materiaal.

Malcolm Chisholm is directeur van Askget.com Inc te New Jersey.

Update

Cordys en Human Inference bieden complete oplossing voor integratie van klantdata

Cordys, leverancier van een geïntegreerd SOA-platform, heeft een wereldwijde samenwerkingsovereenkomst gesloten met Human Inference, leverancier van datakwaliteitssoftware.
De samenwerking verbetert de betrouwbaarheid van klantgegevens door gebruik van een centrale data-hub.
Door deze samenwerking zijn beide partijen in staat hun klanten één oplossing te bieden voor alle aspecten van klantdata-integratie (Customer Data Integration, CDI). Deze kunnen variëren van de kwaliteit van klantgegevens tot de toegankelijkheid van deze informatie binnen een onderneming.
Met behulp van CDI-oplossingen kunnen bedrijven klantinformatie afkomstig uit diverse operationele bronnen samenbrengen in een eenduidig klantbeeld, waardoor zij meer klantgericht kunnen handelen. Datakwaliteit is een absolute voorwaarde voor het creëren en onderhouden van een eenduidig klantbeeld en het voldoen aan wet- en regelgeving voor compliance. Daarnaast is datakwaliteit essentieel voor efficiënte zoekprocessen en betrouwbaardere fraudedetectie.
Zie www.cordys.com en www.humaninference.com

Cognos Go! breidt BI-speelveld uit naar SAP

Cognos, aanbieder in oplossingen voor Business Intelligence en Performance Management, introduceert Cognos Go! Search Service Support for SAP. Met deze nieuwe service kunnen bedrijven die gebruikmaken van Cognos Go! nu ook zoeken naar BI-content in bronnen als mySAP (inclusief SAP-applicaties en datawarehouses). Cognos Go! Search Service support for SAP is per direct beschikbaar.
Cognos Go! voorziet gebruikers van relevante, strategische bedrijfsinformatie via een handige, webgebaseerde zoekfunctie. Door deze zoekfunctie naar SAP-databronnen uit te breiden, kunnen bedrijven transactionele en ongestructureerde SAP- en niet-SAP-data samenbrengen in één geconsolideerd overzicht van performance-informatie, inclusief dashboards, rapporten, analyses, meetgegevens en events. Het zoeken naar informatie en het opstellen van rapporten kost hierdoor minder tijd en moeite, terwijl bestaande BI-investeringen beter benut worden. Cognos Go! biedt SAP-klanten bovendien snel en simpel toegang tot de meest relevante data, dankzij: snelle, complete en relevante resultaten; flexibele toegang; meer zelfbediening; consistente beveiliging; onafhankelijke en soepele ERP-ondersteuning.
Meer informatie: www.cognos.nl