

# Datamodeling

## Wat is het precies en wat kan je ermee?

door: Ed Lute

Zoals vrijwel ieder vakgebied staat ook de ICT bol van de terminologie. Voor insiders zijn de meeste termen en afkortingen wel gesneden koek. Toch blijven de meer specialistische onderwerpen ook voor veel ict'ers veelal steken bij een vage notie. Zo van: de klok wel horen luiden, maar waar die klepel hangt...? Wat te denken van het begrip: datamodeling? Een specialist op het gebied van Business Intelligence of dataprocessing, zal zonder twijfel steekhoudend kunnen toelichten wat deze term behelst. Voor al die anderen zijn we eens de diepte in gedoken.

Business Intelligence (BI) is al geruime tijd een belangrijke discipline binnen de ICT. De groeiende noodzaak van bedrijven om steeds sneller in te spelen op marktveranderingen, dwingt bedrijven om BI echt serieus te nemen. Lange tijd werd verondersteld dat als je maar zoveel mogelijk data verzamelde, je dan marktbevingen tot in details in kaart kon brengen. Helaas ligt dat iets gecompliceerder en moeten de grote hoeveelheden ruwe data nog door een stevig proces heen, voordat ze überhaupt tot bruikbare rapportages kunnen worden omgevormd. Inmiddels is dat wel bekend, maar dan nog zijn er echte BI-specialisten nodig om al die data te modelleren zodat er bruikbare informatie uit kan worden gedestilleerd.

### BI-volwassen

Er bestaan diverse methoden over datamodeling en het opzetten van een datawarehouse. De bekendste zijn wel de methode van Ralph Kimball en Data Vault van Dan Linstedt. Vanzelfsprekend geldt een aantal kenmerken voor alle datamodeling, maar er bestaat niet zoiets als één universele aanpak die in iedere situatie tot optimale resultaten leidt. Veel van die aanpak hangt samen met hoe BI-volwassen een partij is en zaken als datakwaliteit en de wensen van de business. Uiteindelijk moet het zo zijn dat een organisatie met een specifieke informatiebehoefte de beste combinatie van producten en expertise krijgt.

Dit artikel kwam mede tot stand op basis van een interview met Lodewijk Wiggers, partner bij Ensior en daar verantwoordelijk voor oplossingen en techniek.

### Informatiebehoefte

Datamodeling speelt zich vaak af aan het begin van een BI-traject. De weg naar een rapport begint meestal met het modelleren van de ruwe data. In wezen is dat het rangschikken van data op een andere plek, zodanig dat je er handiger mee kunt werken. Aan de informatie die in de data besloten ligt, verander je niets tijdens dit traject. Je gooit niets weg, je rangschikt het alleen op een andere manier. In een financiële omgeving wordt bijvoorbeeld gewerkt met zogeheten 'actuals' – zoals omzetgegevens van de afgelopen dagen – en met financiële informatie omtrent budget of planning. Deze informatie zit in de bronsystemen meestal niet in dezelfde tabel, maar de manager wil die informatie bij elkaar zien, om gemakkelijker bepaalde verbanden te kunnen leggen. Vanuit het oogpunt van rapportage zet je dus bij elkaar wat vanuit die informatiebehoefte bij elkaar hoort te staan. Bronsystemen kunnen in het meest complexe geval 25000 tabellen bevatten. Maar zelfs als dat er 'slechts' 3000 zijn, verhindert de complexiteit het samenstellen van begrijpelijke en bruikbare rapportage. Daarom bekijk je bij de eerste stap van datamodeling welke broninformatie interessant is. Dit leidt tot een subset van tabellen. Je simplificeert zogezegd dat hele grote, complexe model van vele duizenden tabellen naar de essentie van wat je nodig hebt. Binnen die subset van tabellen wordt bekeken hoe nu de werkelijke relatie van de verschillende informatiedelen in elkaar zit. Hiermee bedoelen we niet de relaties zoals die in het data-



model van de leverancier zijn vastgelegd, maar hoe het bij de klant is geconfigureerd en hoe het is gevuld. Dit omvormen en versimpelen tot een beter hanteerbaar model kan tot op zekere hoogte worden geautomatiseerd.

### Profiling

Laten we met dit proces in het achterhoofd, nog een stapje verder de diepte ingaan. Modelleren is dus eerst focussen op: wat heb ik nodig? Dit proces van het kaf van het koren scheiden, heet profiling. Van de 25000 tabellen die SAP kan bevatten, kunnen met een geautomatiseerde oplossing de 400 tot 500 echt relevante tabellen worden geselecteerd. Deze tabellen bevatten de userdata. Maar zelfs die 450 tabellen zijn nog veel te onoverzichtelijk. Je gaat

daarom kijken naar hoe data aan elkaar is gerelateerd. Ieder datamodel laat zogeheten meer-op-meer relaties toe. Je hebt binnen het model een order en je hebt orderregels, wat één-op-meer wordt genoemd. Maar soms is er ook sprake van meer-op-meer relaties. Stel dat een bedrijf van iedere klant één factuuradres heeft, maar dat het model zo is geconfigureerd dat er meer factuuradressen kunnen worden ingevoerd. De data wijzen echter uit dat er in alle gevallen maar één factuuradres is ingevoerd. In dat geval kun je het model voor die 450 tabellen versimpelen door een meer-op-meer relatie terug te brengen naar een één-op-meer relatie. Op basis van wat de data uitwijzen ben je dus aan het modelleren. Met dit zogeheten datadriven modelleren dik je steeds verder in tot de essentie van de data.

### Kardinaliteit

Als het proces van dataprofiling is afgerond, wordt er gekeken naar relaties tussen dingen. De gangbare term voor deze stap is het onderzoek naar de kardinaliteit. De kardinaliteit in een relatie zegt iets over de hoeveelheid van de ene tabel ten opzichte van de hoeveelheid van de andere tabel. We spraken eerder reeds over orders en orderregels. Het komt voor dat een organisatie in haar bedrijfsvoering altijd maar één orderregel gebruikt, omdat de verkoop bijvoorbeeld altijd uit slechts één artikel bestaat. Het datamodel van de leverancier biedt dan wel de mogelijkheid van één-op-meer-relaties, maar voor deze organisatie is het effectief altijd één-op-één. Dat is de kardinaliteit die onderzocht moet worden om daarmee het model nog verder te versimpelen.

### Pivoting

Een ander belangrijk onderdeel van datamodeling is het zogeheten pivoting. Veel applicaties werken met korte lijstjes, zoals klantgroepen. Zo werkt een bankinstelling voor de publieke sector met klantgroepen als waterschappen, gemeenten, provincies, woningcorporaties, zorginstellingen enzovoort. Andere organisaties werken met nog veel meer klantgroepen. Sommige klanten kunnen onder meerdere klantgroepen vallen. Stel nu dat klanten in vier groepen kunnen worden ingedeeld, terwijl er misschien in totaal wel twintig klantgroepen bestaan. In dat geval ontstaan er diverse combinaties van die groepen. Iemand behoort maximaal tot vier groepen, dus daar zet je dan vier kruisjes bij. In applicaties wordt een dergelijke indeling vaak onderhouden met relatietabellen. Dus klanten die bij die ene groep horen, krijgen er een record bij. De business wil graag in één oogopslag zien bij welke klantgroepen een klant hoort. Het is veel te omslachtig wanneer een manager uit die wirwar van klantgroepen eerst de kruisjes bij elkaar moet zoeken. Daarom klap je het in zeker zin om, letterlijk: pivoting. Wat eerst regels waren, worden na het pivotten kolommen. Je weet namelijk dat iedere klant tot maximaal vier klantgroepen kan behoren. Door nu te zorgen dat de aangekruiste regels in kolommen worden getoond, kun je een veel overzichtelijker tabel laten zien. Pivoting is eigenlijk niets anders dan de informatie omzetten naar een andere indeling. Dit is een heel krachtig middel om het model weer verder te versimpelen. Alleen werkt het niet als de ordergrootte boven de twintig

Vaak weet de business niet zo goed wat ze nu precies willen zien

komt. Worden het er meer dan is pivoting eigenlijk niet meer realiseerbaar en leidt het eerder tot een complexer geheel dan dat het simplificeert.

### Datakwaliteit

Vanzelfsprekend speelt de kwaliteit van de data door dit hele proces heen een belangrijke rol. Zo kan er bijvoorbeeld 'ruis' zitten in een applicatie. Daarvan spreek je wanneer relaties voor 99,9 procent identiek lopen, maar dat 0,1 procent daarvan afwijkt. Dit kan bijvoorbeeld een erfenis zijn van de testfase van een applicatie, waarbij er iets werd getest, wat vervolgens na de inrichting van de applicatie nooit meer zo is gebruikt. Het is dan weliswaar niet belangrijke ruis, maar het moet wel onderzocht worden. Meestal komen dergelijke uitzonderingen naar voren tijdens de diverse fasen van datamodeling. Het heeft zelden zin om op voorhand iets te willen doen aan de datakwaliteit. Na het proces van datadriven modeling hebben de experts een lijstje met vragen en opmerkingen over zaken die hen zijn opgevallen.

### Overleg met de business

Met profiling en datadriven modeling kom je al snel tot een 0.1-versie van je datamodel. Tot dan toe is er nog nauwelijks gekeken naar wat er binnen het proces gebeurt, en is er ook nog geen overleg met de business geweest. Het is een algemene misvatting dat er eerst uitvoerig met de business moet worden overlegd om überhaupt tot een goed model te kunnen komen. Deze indruk leeft met name bij organisaties die nog niet zo BI-volwassen zijn. Zij menen dat de business vanaf de eerste stap aan de experts heel veel input moet geven over hun processen om tot een goed datamodel te komen. Maar in deze fase maakt het eigenlijk niet uit of het tabel A, B of C is of dat wat in de data als 'fabriek' staat vermeld in werkelijkheid 'vestiging' moet heten. Pas wanneer de finetuning in beeld komt en de specifieke wensen van de business moeten worden vormgegeven, is het moment aangebroken om met elkaar in overleg te treden. Sterker nog, eerder overleg met de business kan tot nodeloze verwarring leiden, omdat de gebruikers een bepaald idee hebben over hoe de data in elkaar zit, wat niet altijd met de werkelijkheid overeenkomt.

### Grote en kleine stappen

De tijd die nodig is om te komen tot een 0.1-versie van je datamodel, hangt uiteraard af van de complexiteit van de bron. Een groot voordeel hierbij is dat dit proces voor een belangrijk deel te automatiseren is. Daarom kun je eerder spreken van enkele dagen dan van enkele weken om te komen tot een werkbaar model. Dit is vooral werkbaar voor

de verdere modellering. Voor de business zelf is het dan meestal nog niet bruikbaar. Dat komt omdat de business vaak niet zo goed weet wat ze nu precies willen zien. Die versie 0.1 wordt dan ook doorgaans gebruikt om de discussie met de business te starten, waarna er een verfijning van dat model kan worden gemaakt, naar versie 0.2. Dus eerst grote stappen ernaartoe, en tussendoor kleine stapjes om te verfijnen.

### Rapportagebehoefte

De drie stappen: 1) profiling, 2) onderzoek naar de kardinaliteit en 3) pivoting vormen tezamen het kernproces van datamodeling. Het hele proces komt neer op het simplificeren van het datamodel zodat het in de vervolgstappen snel en effectief kan worden afgestemd op de informatievraag van de business. Tot hier toe is er nog niet gekeken naar wat de business wel of niet interessant vindt. Er is alleen gekeken naar de data zelf die zegt dat ze wel of niet interessant zijn. Hierna moet het zo zijn dat alles wat de gebruikers op schermen of rapporten aanwijzen, terug te vinden moet zijn in het datamodel. In deze fase vindt de afstemming plaats of de manier waarop de data in tabellen en combinaties zijn opgeslagen, het meest handig is voor de business. Wanneer je bijvoorbeeld iedere keer over vijf tabellen heen moet samenvoegen om tot een antwoord te komen voor bijna alle rapporten, is dat blijkbaar niet de beste opslagmethode voor die informatie. Met deze bevindingen wordt het datamodel gesimplificeerd met de rapportagebehoefte als uitgangspunt.

### Finetuning

In deze fase kijk je naar de KPI's, de definities en de formules. Waar leg je de formules vast? Doe je dat in het rapport, in de metadatalaag of in de database zelf? Afhankelijk van of het om een generiek verzoek om rapportage gaat of een specifiek verzoek, moet je de positie van de formules kiezen. Dit is de echte finetuning. De stappen naar de 0.1-versie maken meestal het leeuwendeel uit van het hele proces van ruwe data naar rapportage, en de beste resultaten worden toch bereikt als dit in de handen van specialisten wordt gegeven. Uiteindelijk gaat een organisatie deze exercitie immers aan om de eigen concurrentiepositie te versterken. En zonder datamodeling krijg je met de huidige datavolumes zelden een overzicht van je kerninformatie. Bronapplicaties zijn namelijk ongeschikt om er rechtstreeks een informatievraag op af te vuren. Door een model te creëren waarin die bronnen worden gerepresenteerd, kun je komen tot zinvolle rapportages.

### Versimpeling

Met profiling kun je veel tabellen en kolommen die ogenschijnlijk vulling bevatten, maar dit in werkelijkheid niet doen, verwijderen. Dit klinkt alsof je dan het risico loopt dat je data kwijtraakt die je misschien later nog nodig zou kunnen hebben. Dat is niet het geval, want het enige dat je weggooit, is datavervuiling. De ervaring leert dat van de 250 kolommen waaruit een tabel kan bestaan, er hooguit 40 echte userinformatie bevatten. Veel kolommen zijn echt leeg, andere bevatten uitsluitend dezelfde, constante waarde. Een kolom die zonder uitzondering dezelfde waarde bevat, is geen zinvolle informatie voor je rapportage. Na deze geautomatiseerde versimpeling bevat het datamodel alle mogelijke relevante userdata. Het is dus niet meer zo – zoals traditioneel gebruikelijk is – dat alleen die data in het model worden opgenomen die door de business zijn aangewezen. Waar het dan om gaat is om er zeker van te zijn dat de diverse data-elementen op de handigste plek staan voor de business om rapporten te maken.

Geef alles uit handen, behalve de touwtjes

De PrioX Service Management Suite biedt grip op de uitbesteding van diensten. Een eenvoudig te bedienen dashboard en heldere rapportages leveren een onafhankelijke blik op de kwaliteit van alle interne en externe dienstverleners.

**priox**  
service management solutions

www.priox.com +31 (0)3 18 21 01 01 info@priox.com

ENJOY 1 MILLIMETER ATTRACTION

WORK BETTER

CHOOSE EIZO

De EV2455 met sensationeel smalle randen voor grenzeloze beeldkwaliteit. Kies voor ons nieuwe 24 inch model en geniet van de smalste schermranden die EIZO ooit heeft ontwikkeld. Met maar 1 mm breedte is dat zelfs voor een vlo minuscule. En biedt veel ruimte voor de beste beeldkwaliteit en innovatieve EcoView functies. www.eizo.nl/ultraslim

**EIZO**